Volume 82, Number 5, September/October 1974

Martin Feldstein: Social Security, Induced Retirement, and Aggregate Capital Accumulation

William H. Oakland: Public Goods, Perfect Competition and Underproduction

Matthew Edel and Elliott Sclar: Taxes, Spending, and Property Values: Supply Adjustment in a Tiebout-Oates Model

Wolfgang Mayer: Short-Run and Long-Run Equilibrium for a Small Open Economy

Arthur E. Gandolfi: Stability of the Demand for Money during the Great Contraction—1929–1933

Richard Layard and George Psacharopoulos: The Screening Hypothesis and the Returns to Education

Ben Branch: Research and Development Activity and Profitability: A Distributed Lag Analysis

The University of Chicago Press

JOURNAL OF
POLITICAL ECONOMY

Edited by
ROBERT J. BARRO HARRY G. JOHNSON
SAM PELTZMAN GEORGE J. STIGLER
In cooperation with OTHER MEMBERS of the DEPARTMENT OF
ECONOMICS and the GRADUATE SCHOOL OF BUSINESS
of the UNIVERSITY OF CHICAGO
AND OUTSIDE REFEREES
(SEE INSIDE BACK COVER)
DIANE WELCH, Editorial Assistant

The Journal of Political Economy is published bimonthly in February, April, June, August, October, and December by The University of Chicago Press. Subscription rates, U.S.A.: institutions, 1 year $20.00, 2 years $38.00, 3 years $55.00; individuals, 1 year $15.00, 2 years $28.00, 3 years $40.00. Student subscription rate, U.S.A.: 1 year $9.00 (letter from professor must accompany subscription). Other countries add $1.00 for each year's subscription to cover postage. Single copy rates: institutions $4.00, individuals $3.00. Supplements: institutions $4.50, individuals $3.50. Back issues are available from 1962 (vol. 70). Make all remittances payable to *Journal of Political Economy*, The University of Chicago Press, in United States currency or its equivalent.

Claims for missing numbers should be made within the month following the regular month of publication. The publishers expect to supply missing numbers free only when losses have been sustained in transit and when the reserve stock will permit.

Business correspondence should be addressed to The University of Chicago Press, 5801 Ellis Avenue, Chicago, Illinois 60637.

Letters to the editors and manuscripts should be addressed to the Editor of the *Journal of Political Economy*, 1126 East 59th Street, Chicago, Illinois 60637. **Manuscripts should be submitted in duplicate (3 copies if possible), accompanied by a $25.00 submission fee made payable to the Journal.** The fee will be refunded to authors whose papers are accepted for publication in the *Journal* The proceeds from the submission fees are used to pay for refereeing services. Accepted manuscripts must be typed according to the University of Chicago *Manual of Style*. References should be typed double-spaced at the end of the article. Footnotes should be numbered in sequence and double-spaced following the references. Tables should follow the footnotes. Originals of the figures, drawn in india ink, should be submitted if the manuscript is accepted. Abstracts not exceeding 100 words should be submitted in duplicate along with the manuscript.

Applications for permission to quote from this *Journal* should be addressed to The University of Chicago Press and will be freely granted. Reproduction of articles requires written permission from the publisher and the author.

Reprinted volumes 1-72 available from Walter J. Johnson, Inc., 111 Fifth Avenue, New York, New York 10003. Volumes available **in microfilm** from University Microfilms, 300 North Zeeb Road, Ann Arbor, Michigan 48106; **in microfiche** from J. S. Canner & Co., 49-65 Lansdowne Street, Boston, Massachusetts 02215.

Notice to subscribers: If you change your address, please notify us and your local postmaster immediately, giving *both* your old and your new address. *Allow four weeks for the change.*

Second-class postage paid at Chicago, Illinois, and at additional mailing office.
© 1974 by The University of Chicago. All rights reserved.
ISSN 0022-3808
PRINTED IN U.S.A.

Journal of Political Economy

EDITED BY ROBERT J. BARRO, HARRY G. JOHNSON,
SAM PELTZMAN, AND GEORGE J. STIGLER
IN CO-OPERATION WITH OTHER MEMBERS OF THE
DEPARTMENT OF ECONOMICS AND THE
GRADUATE SCHOOL OF BUSINESS OF
THE UNIVERSITY OF CHICAGO
AND OUTSIDE REFEREES

Volume 82

January/February–November/December 1974

THE UNIVERSITY OF CHICAGO PRESS
CHICAGO, ILLINOIS

© 1974 BY THE UNIVERSITY OF CHICAGO. ALL RIGHTS RESERVED.
PUBLISHED JANUARY/FEBRUARY,
MARCH/APRIL, MAY/JUNE, JULY/AUGUST, SEPTEMBER/OCTOBER,
NOVEMBER/DECEMBER 1974
BY THE UNIVERSITY OF CHICAGO PRESS

Journal of Political Economy

Volume 82 Number 5 September/October 1974

Articles

905 Social Security, Induced Retirement, and Aggregate Capital Accumulation
Martin Feldstein

927 Public Goods, Perfect Competition, and Underproduction
William H. Oakland

941 Taxes, Spending, and Property Values: Supply Adjustment in a Tiebout-Oates Model
Matthew Edel and Elliott Sclar

955 Short-Run and Long-Run Equilibrium for a Small Open Economy
Wolfgang Mayer

969 Stability of the Demand for Money during the Great Contraction—1929–1933
Arthur E. Gandolfi

985 The Screening Hypothesis and the Returns to Education
Richard Layard and George Psacharopoulos

999 Research and Development Activity and Profitability: Distributed Lag Analysis
Ben Branch

Communications

1013 A Note on Border-Tax Adjustments
James E. Meade

1017 Policy Influence without Policy Choice
Donald C. Cell

1027 Time-Series Changes in Personal Income Inequality in the United States from 1939 with Projections to 1985: A Comment
Aba Schwartz and Donald O. Parsons

1033 Time-Series Changes in Personal Income Inequality in the United States from 1939 with Projections to 1985: Reply
Barry Chiswick and Jacob Mincer

1035 **Market Allocation of Exhaustive Resources**
Oliver Scott Goldsmith

1041 How Analysis Affects Decision-Makers' Views: A Bayesian Model
Rolla Edward Park

1049 On Some Unhappy Consequences of Professor Liebhafsky's Consumers' Demand for Psychiatric Help
Jean-Claude Koeune

1053 **Book Reviews**

Orley Ashenfelter and Albert Rees, eds., *Discrimination in Labor Markets*
Ronald G. Ehrenberg

Kenneth E. Boulding, *The Economy of Love and Fear: A Preface to Grants Economics*
Harry G. Johnson

Harry A. Cahill, *The China Trade and U.S. Tariffs*
Robert F. Dernberger

Richard E. Caves and Ronald W. Jones, *World Trade and Payments: An Introduction*
Anne O. Krueger

Arnold R. Weber, *In Pursuit of Price Stability: The Wage-Price Freeze of 1971*
William Poole

Journal of Political Economy

Volume 82 Number 6 November/December 1974

Articles

1063 A Theory of Social Interactions
Gary S. Becker

1095 Are Government Bonds Net Wealth?
Robert J. Barro

1119 Wage Comparisons—a Selectivity Bias
Reuben Gronau

1145 Comments on Selectivity Biases in Wage Comparisons
H. Gregg Lewis

1157 Progressivity and Investment Yields under U.S. Income Taxation
Martin J. Bailey

1177 Adam Smith on Profits—Paradox Lost and Regained
Nathan Rosenberg

1191 Tariffs and the Distribution of Income: The Importance of Fa
Specificity, Substitutability, and Intensity in the Short and L
Run
Michael Mussa

1205 The Stability of the Demand-for-Money Function in the United States 1901–1965
Mohsin S. Khan

1221 Estimating the Long-Run Demand for Money from Time-Series Data
Rodney L. Jacobs

Communications

1239 External Economies and the Production Frontier
J. R. Gould

1249 The Staple Theory and Its Empirical Measurement
Dwight Grant

1255 Risk, Job Search, and Income Distribution
Christopher A. Pissarides

The Role of Absolute and Relative Wage Differentials in the Choice of a Work Activity
Shmuel Sharir and Yoram Weiss

Stability in the Pareto Law and the Popular Folklore: A Comment
Daniel R. Vining, Jr.

The Real Stock of Money and Velocity in a Wartime Economy: A Comment
Joong-Koon Lee

The Real Stock of Money and Velocity in a Wartime Economy: A Reply
G. E. Makinen

Miscellany

What Is the Optimal Level of Pollution?
E. J. Mishan

Review Article

Poor Old Capitalism: A Review Article
Evsey D. Domar

Book Reviews

Anthony Downs, *Opening Up the Suburbs*
Edwin S. Mills
Ronald G. Ehrenberg, *The Demand for State and Local Government Employees*
Richard W. Parks

Index to Volume 82

Social Security, Induced Retirement, and Aggregate Capital Accumulation

---

Martin Feldstein

Harvard University

For the great majority of Americans, the most important form of household wealth is the anticipated social security retirement benefits. In 1971 the aggregate value of these annuities was approximately $2 trillion or some 60 percent of other household assets. This paper uses an extended life-cycle model to analyze the impact of social security on the individual's simultaneous decision about retirement and saving. Econometric evidence, using an estimated time series of "social security wealth," indicates that social security depresses personal saving by 30–50 percent. Implications of this research for the theory of consumption and for the level and distribution of income are discussed.

For the great majority of Americans, the most important form of house wealth is the anticipated social security retirement benefits. In 1971 aggregate value of these annuities was approximately $2 trillion or 60 percent of other household assets. Neither the theoretical nor empirical analysis of household consumption behavior has given adequate attention to the existence and growth of social security.

This paper shows that the effect of social security is very important and is more complex than previous discussions have recognized. Section 1 presents a theoretical analysis that emphasizes the impact of social security on the individual's simultaneous decision about retirement and saving. In contrast to previous models of life-cycle saving, the extent of retirement is endogenous. Section 2 discusses the use of an aggregate consumption

This paper is part of a larger study of the effects of fiscal policies on capital formation and income distribution. I am grateful to the National Science Foundation for financial support, and to Roger Gordon, Alicia Munnell, and Anthony Pellechio for assistance with this research. I have also benefited from discussions and comments on a previous draft by Henry Aaron, John Flemming, Milton Friedman, Thomas Mayer, Franco Modigliani, Richard Musgrave, and the referees of this *Journal*.

[*Journal of Political Economy*, 1974, vol. 82, no. 5]

© 1974 by The University of Chicago. All rights reserved.

unction to assess the effect of social security. A key variable in the analysis an estimated time series of "social security wealth," that is, the present alue of the social security annuities to which individuals are entitled. The data sources are described in Section 3, and the parameter estimates re presented in Section 4. The evidence indicates that social security has ubstantially decreased aggregate capital accumulation; specific estimates re presented in Section 5. A final section discusses some of the impli- ations for the level and distribution of income.

A Life-Cycle Model with Induced Retirement

ver since Harrod's (1948) discussion of "hump saving," economists ve recognized the importance of saving during working years for nsumption during retirement.[1] Although there are a variety of other otives for saving, the life-cycle hypothesis provides a particularly itable framework for discussing the effects of social security. The most vious implication of this familiar model is that social security, by oviding income during retirement, reduces the amount of saving ring the working years. More specifically, if the combination of social urity tax and benefits has no net income effect, that is, if the indi- ual's lifetime budget constraint is unchanged, savings will be reduced just enough to leave consumption during retirement unchanged.

The possible importance of social security has been recognized but never been formally incorporated into the theoretical or empirical lysis of the life-cycle model. Friedman (1957, p. 123) noted that social rity "would clearly tend to reduce the need for private reserves and reduce private savings" but made no allowance for this in his analysis oth aggregate and cross-section savings. The potential effect of social rity has also been ignored in Modigliani's own tests of the life-cycle othesis using time-series data for the United States (Ando and ligliani 1963) and aggregate data for a cross-section of countries digliani 1970). Mayer's (1972) recent book provides an extensive w of previous studies of the life-cycle hypothesis but no examples dies in which the effect of social security had been considered.[2]

Harrod explicitly recognized, Fisher (1930) provided the basic model of con- ion behavior from which his own analysis is derived. Harrod's contribution was to Fisher's two-period model to the stages of the life cycle and to explore some of its ations for Keynesian macroeconomics. The Modigliani-Brumberg (1954) paper ed this to a more general multiperiod analysis and derived the implications for the of wealth on consumption.

a cross-section study of savings in different countries, Aaron (1967) found evidence gher social security benefits decreased private saving. His basic equation is in- nt with the life-cycle hypothesis, ignoring variables such as the growth rates of and of population that would be important determinants of intercountry differ- savings (see Modigliani 1970). This may explain why Aaron later reported that, In contrast to the implications of the life-cycle theory, students of social security have generally argued that social security is not likely to have a substantial effect on personal savings and that it might even cause such savings to increase.[3] This conclusion is based largely on the evidence of Katona (1964) and Cagan (1965) that persons covered by private pensions do not save less and may save more than those persons not covered by pensions. More specifically, Cagan analyzed data generated by a mail survey of *Consumer Reports* subscribers and found that the average savings rate was slightly higher for those with pensions than for those without and was also higher for those with vested pensions than for those whose pensions were not vested. A regression equation implied that an increase in the individual's rate of pension contribution was associated with a higher level of direct ("discretionary") saving. Katona analyzed data collected by a University of Michigan survey of randomly selected households and also found that participation in a pension plan raised savings rates when age and current income were taken into account.

Cagan explained his surprising results in terms of a "recognition effect": when an individual is forced to participate in a pension plan, he recognizes for the first time the importance of saving for his old age. Participation in a pension plan has an educational effect; more formally, it changes the individual's utility function as he perceives it ex ante during his working years. Katona added to this a second explanation: the "goal gradient" hypothesis borrowed from psychological res on the forming of aspirations. According to this theory, "effort is int the closer one is to one's goal" (Katona 1964, p. 4). In more conven economic terms, this would imply that individual preferences are selves a function of the opportunity set or of the initial position, a dr departure from the usual assumption of economic analysis.

The findings of Cagan and Katona can be explained without invoking a recognition effect or a model of changing preferences by extending the life-cycle model to make the extent of retirement endogenous. Workers who are covered by pensions have an incentive to retire earlier than they otherwise would. To receive a pension requires retiring from a current job and generally involves loss of union seniority. Even if the individual is permitted to draw the pension after taking a new job, the loss of

---

when his equation was reestimated with data for a different year, the earlier effect of social security was no longer significant (Aaron 1968). A more recent international study using an extended life-cycle model to assess the impact of social security found that higher social security benefits and broader coverage do substantially reduce private savings (see Feldstein 1974).

[3] See, e.g., Pechman, Aaron, and Taussig (1968). Ever since Harris's (1941) study, estimates of the effect of social security on savings have concentrated on the impact of the reduction in disposable income (because of the payroll tax) rather than the life-cycle wealth effect of social security. For a more recent example of this, see Tax Foundation (1967).

seniority and of job-specific skills generally entails a substantial fall in the available wage. The pension therefore acts as a combination of an annual lump sum grant and a tax on earnings after the standard retirement age. The result is to reduce the labor supply of pension recipients, generally through earlier retirement. The pension, therefore has two effects on personal savings: (1) it reduces personal saving because it substitutes for household assets, but (2) it also increases personal saving because it lengthens the period of retirement over which accumulated assets will be spread. The net effect of the pension depends on the relative strength of these two forces.

Social security has a similar dual impact. The "tax" on earnings after age 65 is higher and even more obvious than with private pensions. The "earnings test" in the current social security law provides that a potential recipient loses his social security benefits if he earns more than $2,400 per year, thus requiring retirement from regular employment as a condition of receiving benefits. For an individual who in the absence of social security would have retired at age 65 to consume the income and principal of his accumulated assets, the social security benefits have the unambiguous effect of reducing saving. For those who would otherwise have worked beyond age 65, social security would generally (but not always) induce retirement at an earlier age. For such individuals, the effect of social security on savings is uncertain. In the extreme case of the individual who planned to work as long as he was able and then to be supported by his children or at public expense, the inducement to plan an early retirement could only increase savings. Although additional cases could be distinguished, it is already clear that the two countervailing effects make the net impact of social security ambiguous.[4]

The dual effect of social security is illustrated in figure 1. The horizontal axis measures income and consumption before age 65 (Y_1 and C_1), and the vertical axis measures income and consumption after that age (Y_2 and C_2). For simplicity, pretax income before age 65 is assumed to be unaffected by the introduction of social security. Consider first an individual who in the absence of social security would be fully retired after age 65. His earnings are described by point A: Y_{1A} in period 1 but zero in period 2. Faced with an opportunity to save at the rate of interest implied by the budget line through point A, the individual chooses the pair of consumption levels denoted by point I. During his preretirement period, he

---

[4] The introduction of social security also had the effect of removing a peculiar "tax" on savings by low-income individuals. Before social security, an individual could generally receive support at public expense only after he had exhausted all of his own assets. For the individual whose income was so low that a reasonable rate of saving during working years permitted a level of retirement consumption little better than that provided by public assistance, there would be a strong incentive to save nothing. In the same way, aged persons who lived with their children might find that their own assets, if any, did not alter retirement conditions and, therefore, were in effect taxed away.

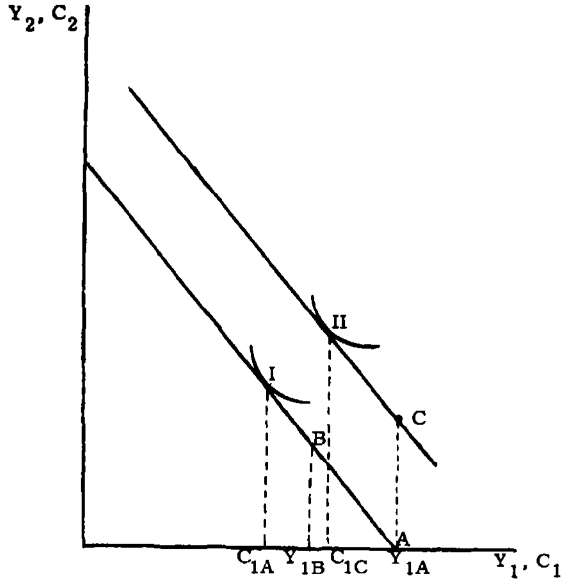


FIG. 1

consumes C_{1A} and saves $Y_{1A} - C_{1A}$. Point B describes that indivi initial position after the introduction of a social security tax; his s period income is raised by the amount of the benefit, and his first-p income is reduced by the payroll tax to Y_{1B}; the difference betwee and Y_{1B} is just sufficient to finance the benefits at the market r interest. Since the introduction of social security does not alte individual's budget line, he retains the original equilibrium consumption point (I). The reduction in disposable income implies that saving is reduced from $Y_{1A} - C_{1A}$ to $Y_{1B} - C_{1A}$, that is, by the amount of the tax ($Y_{1A} - Y_{1B}$).

This reduction in saving can be contrasted with the experience of an individual who in the absence of social security would continue to work for some time after age 65. Point C indicates an initial position with positive earnings in the second period and the same level of first-period earnings as point A. The new equilibrium consumption pair is denoted by point II; first-period consumption is C_{1C}, and the corresponding saving is the relatively small amount $Y_{1A} - C_{1C}$. If the social security program induces him to retire completely at age 65, his initial position shifts from C to B, that is, his period 1 disposable income is reduced by the amount of the payroll tax, while his period 2 disposable income falls because earnings are reduced to zero and not fully replaced by social security benefits. With initial position B, the individual chooses consumption pair I.[5] The effect of introducing social security is, therefore, to change savings from $Y_{1A} - C_{1C}$ to $Y_{1B} - C_{1A}$. The current figure is constructed to make $Y_{1B} - C_{1A} > Y_{1A} - C_{1C}$, that is, the introduction of social security increases saving. Alternative assumptions about the consumption-expansion path could yield different results.

These two examples are sufficient to illustrate how social security will increase savings by some and decrease savings by others. A more general mathematical analysis of the effect of social security in the extended life-cycle model shows[6] that, *if the labor supply is fixed in both periods*, the introduction of a social security program reduces personal saving even if there are positive second-period earnings. When the labor supply in the second period is endogenous, social security benefits reduce second-period labor (i.e., induce earlier retirement), but the effect on savings is ambiguous.

As is so often the case, a theoretical analysis can illuminate the ways in which a public policy affects individual behavior, but it cannot yield an estimate of the magnitude of the effect nor even an unambiguous conclusion about its sign. For this we must turn to an empirical investigation. A full study of this issue would require examining the effect of social security on retirement plans in different circumstances and of the dependence of individual savings on both the retirement plans and the expected social security benefits. The task of the current study is more modest: to assess how the introduction and growth of social security have affected aggregate personal saving and the national accumulation of capital.

II. An Aggregate Consumption Function with Social Security Wealth

To assess the effect of social security, I have adapted the specification of the consumption function used by Ando and Modigliani (1963) and have added a social security wealth variable. The starting point for the analysis is thus:

$$C_t = \alpha + \beta_1 Y_t + \gamma_1 W_{t-1}, \tag{1}$$

where C_t is consumer expenditure, Y_t is permanent income (specified below as a distributed lag on past values of disposable income),[7] and W_t

---

[5] The indifference curves through points I and II are not comparable. The second individual prefers point I to point II because I is associated with not working after age 65.

[6] This analysis is presented in Feldstein (1974).

[7] Ando and Modigliani originally used *labor* income only but, because of the difficulties of defining and measuring disposable labor income, Modigliani (1971) has more recently abandoned labor income and used total disposable income. Ando and Modigliani also introduced an adjustment for unemployment. Although the use of a permanent income variable can serve this purpose, the unemployment rate will be examined explicitly.

is the stock of household wealth at the end of year t (excluding social security wealth).

Equation 2 introduces the social security wealth variable (SSW_t) into this consumption function:

$$C_t = \alpha + \beta_1 Y_t + \gamma_1 W_{t-1} + \gamma_2 SSW_t. \qquad (2)$$

Two different definitions of SSW are used in the current study. Gross social security wealth ($SSWG_t$) is the present value in year t of the retirement benefits which could eventually be claimed by all those who are either in the labor force or already retired in year t. The calculation of these present values reflects survival probabilities as well as the discounting of future receipts. Net social security wealth ($SSWN_t$) equals gross social security wealth minus the present value of the social security taxes to be paid by those who are currently in the labor force. Net social security wealth is positive because the social security program involves a transfer to those who are currently working or retired from those who are yet to enter the labor force.

Since disposable income already excludes social security taxes, the gross social security wealth variable is probably the correct specification. Using the net social security wealth is equivalent to taking into account not only the current payroll tax but the present value of all future taxes. This is more appropriate if one views the social security program as creating for each individual an amount of wealth equal to the pr value of the difference between his future benefits and his future The consumption function has been estimated with both definit Fortunately, both sets of estimates have the same implications.

The basic logic of the calculations used to evaluate $SSWG$ and SS can be explained briefly.[8] Consider an unmarried worker who is ag in year t. If he remains single and survives to age 65, he will be entitled to an annual social security benefit in the amount $b_{a,t}$. In estimating the average value of future benefits for a single surviving annuitant, it would be wrong to assume that the schedule of benefits provided by law in year t would remain in effect. The history of social security shows continually rising benefit levels, a fact that individuals no doubt perceive when they contemplate the order of magnitude of their own benefits at retirement age. The ratio of annual benefits for retired workers (excluding dependents' benefits) to per capita disposable income has varied without any trend around a mean level of 0.41. The current calculation assumes that $b_{a,t}$ is 0.41 times per capita disposable income in year $t + 65 - a$ when the individual retires, Y_{t+65-a}. Finally, this future value of disposable income is estimated by projecting the current level at a constant rate of growth of disposable per capita income, $Y_{t+65-a} = Y_t(1 + g)^{65-a}$,

---

[8] A full description of the algorithm used is available from the author.

The anticipated benefits at age 65 are thus $b_{a,t} = 0.41\ Y_t(1 + g)^{65-a}$. During his retirement, the annual benefit will continue to grow. For simplicity we assume that this growth also occurs at rate g; thus at age $n > 65$, his annual benefit is $b_{a,t}(1 + g)^{n-65}$.

Let $S_{i,j}$ denote the probability that a man age i survives to at least age j, and let d be the rate at which the individual discounts expected future benefits. At age 65, his social security annuity will have a value of

$$\sum_{n \geq 65} S_{65,n} b_{a,t}(1 + g)^{n-65}(1 + d)^{-(n-65)};$$

for practical purposes the sum may be truncated at age 100. At time t (when the individual is age a) the future annuity has value:

$$A_{a,t} = S_{a,65}(1 + d)^{-(65-a)} \sum_{n \geq 65} S_{65,n} b_{a,t}(1 + g)^{n-65}(1 + d)^{-n+65}.$$

After substituting for $b_{a,t}$ we obtain:

$$\begin{aligned} A_{a,t} &= 0.41\ Y_t S_{a,65}[(1 + g)/(1 + d)]^{65-a} \\ &\times \sum_{n \geq 65} S_{65,n}[(1 + g)/(1 + d)]^{n-65}. \end{aligned} \tag{3}$$

Analogous calculations are done for single women, for working wives, and for married couples in which only the husband works. The special benefit rules for wives and for working women are reflected in these calculations. In particular, the value of the annuity for a married worker includes the benefits paid to his surviving spouse. The value of $SSWG_t$ is the weighted sum of the $A_{a,t}$ values for each type of worker and for each age group, weighting by the numbers of workers of that type and age covered by social security in year t.[9]

The present value of future social security taxes is obtained by a similar method. A worker who contemplates his future social security taxes recognizes that they will rise because of increases in both the tax rate and the level of taxable earnings. For simplicity, we combine these and assume that the individual forecasts the ratio of social security taxes per covered worker to per capita disposable income, θ_t. An average worker who is age a in year t pays a tax of $T_t = \theta_t Y_t$, independent of his age. He expects that at age m he will pay a tax of $T_{t+m-a} = \theta_{t+m-a} Y_t(1 + g)^{m-a}$. The present value of all his future taxes until age 65 is thus:[10]

$$\text{TAX}_{a,t} = \sum_{m=a}^{64} S_{a,m}\theta_{t+m-a} Y_t[(1 + g)/(1 + d)]^{m-a}. \tag{4}$$

---

[9] An adjustment is made for those who are not currently labor-force participants but can be expected to draw social security benefits at some time in the future.

[10] A special adjustment is made for those over age 65 who are still employed.

The value of θ_t has been rising in an irregular pattern since the introduction of social security and in 1971 reached approximately 0.15 for males and 0.07 for females. The current calculations assume that the individuals correctly foresaw the changes in θ_t before 1971 and assumed that after 1971 the θ_t's would remain at the 1971 levels. The total present value of expected future taxes is a weighted sum of the $TAX_{a,t}$ values for different employee types (single men, married men, etc.) and age groups, weighting by the numbers of covered workers in each group in year t. The net social security wealth value in each year is the difference between $SSWG_t$ and this aggregate tax value.

Despite the complexity of these calculations, the values of $SSWG_t$ and $SSWN_t$ do not have the precision of actuarial estimates. For example, there is no adjustment for future marriages and divorces or for remarriages by surviving beneficiaries. But an exact actuarial estimate is not required for the current analysis. The *SSW* variable should reflect the magnitude of the effect of social security to which households respond. The household's implicit evaluation of social security benefits is likely to reflect their perception of the current standard of living among social security annuitants and of the way in which this has changed in the past. Note that households may therefore be able to respond appropriately to changes in social security without being able to calculate or articulate the value of benefits.

The *SSW* variable values the benefits for which individuals are el at age 65: there is no reduction for the individuals who can be exp to continue to work beyond age 65. In this way, the *SSW* variab presents both the "substitute asset" effect that would reduce saving the "inducement to retire" effect that would increase savings. The est of γ_2, therefore, measures the net effect of the introduction and gr of the social security program. It therefore underestimates the effect on saving of the increasing number of individuals who already plan to retire at 65.

The consumption function of equation 2 should be extended to reflect the impact of corporate saving on household consumption. During the period since the introduction of social security in 1937, the growing tax incentive associated with rising marginal personal income tax rates has induced companies to increase the fraction of total earnings that is saved.[11] The traditional specification of equation 1 implies that dividends have a much greater effect on concurrent consumption than retained earnings. This implies that the tax-induced substitution of retained earnings for dividends during the past 30 years has reduced household consumption. I believe that this is an unintended result of specifying consumer behavior without considering the implication of the corporate form of business. The current paper, therefore, relaxes this restrictive assumption and uses a model developed in Feldstein (1973) and Feldstein and Fane (1973) in which retained earnings have a short-run effect on consumption in addition to their long-run wealth effect. More specifically, the basic specification used in this paper is:

---

[11] See Brittain (1966) for evidence that this has occurred. Feldstein (1970) found that British firms also varied their savings rate in response to tax-induced changes in the opportunity cost of retentions in terms of foregone dividends.

$$C_t = \alpha + \beta_1 Y_t + \beta_2 RE_t + \gamma_1 W_{t-1} + \gamma_2 SSW_t. \qquad (5)$$

Separate estimates with the retained earnings variable excluded show that this does not substantially affect the estimated impact of social security wealth.

III. The Data

The estimates presented below are based on aggregate U.S. data for the periods 1929 through 1971, excluding the years 1941 through 1946. Separate analyses for the postwar period, 1947 through 1971, are also presented.

Consumer expenditure (C) and disposable personal income (YD) are the usual national income account values, deflated to constant 1958 dollars and divided by population. The measurement of retained earnings raises a problem because of inadequate data on true economic depreciation. Undistributed profits as reported in the national income accounts are equal to gross profits minus an estimate of corporate capital consumption. Because of the accounting conventions used to calculate capital consumption, this measure of net retained earnings is likely to be an underestimate of true net corporate saving. Gross retained earnings is obviously an overestimate. Separate equations have been estimated with each specification. The similar implications of the results with both measures indicates that further work to improve the measurement of retained earnings would not alter any conclusions. Only the equations with gross retained earnings are reported.

The wealth variable is the per capita net worth of households at market value expressed in 1958 dollars. The series was estimated by Ando and Modigliani (1963) based on Goldsmith's (1956) earlier study and updated for the FRB-MIT econometric model.[12] The variable W_t refers to the wealth at the end of year t.

The basic data required to calculate the social security wealth variable are published in the annual statistical supplements of the *Social Security Bulletin*. Each year the average level of benefits for retired workers is published as well as the average payroll tax per covered employee and

---

[12] I am grateful to Franco Modigliani for making available the unpublished data.

TABLE 1

SOCIAL SECURITY WEALTH
(BILLIONS OF CONSTANT 1971 DOLLARS)

	GROSS WEALTH		NET WEALTH		
YEAR	*SSWG*1 (1)	*SSWG*5 (2)	*SSWN*1 (3)	*SSWN*5 (4)	GNP (5)
1940	235	70	145	25	321
1950	442	153	227	36	503
1955	690	270	366	81	620
1960	917	380	493	122	691
1965	1,397	596	805	234	875
1971	2,029	875	1,162	342	1,050

the number of covered employees and annuitants by age and demographic group (single men, etc.). Separate survival probabilities ($S_{t,j}$'s) for men and women are obtained from the life tables published by the U.S. Department of Health, Education and Welfare (1968).

The formula for *SSW* also contains the rate of growth of real per capita disposable income (g) and the individual's rate of discount for future real income (d). For the 35 years from the introduction of the social security program to 1972, real per capita disposable income grew a rate of 2 percent per year. The value of d can be considered the ave real rate of interest after tax that was available to savers during the pe since the introduction of social security. The nominal rate of interest Moody's Baa bonds averaged 5.0 percent from 1937 to 1972. The ave inflation of consumer prices was 3.0 percent. The real yield before t was, therefore, 2.0 percent. The yield on bank deposits and savings bonds was lower. Although the return on common stocks and private homes was much higher, investment in common stocks has been relatively unimportant for low- and moderate-income households, and the scope for additional saving in owner-occupied housing is obviously limited. A value of $d = 0.03$, therefore, appears to be an appropriate estimate. Equations 6 and 7 show that g and d always enter together in the form $(1 + d)/(1 + g)$. What matters, therefore, is only this net discount ratio and not the individual values of g and d. The values $g = 0.02$ and $d = 0.03$ imply a ratio of 1.01. To test the sensitivity of the regression analysis to this value, estimates were also made with *SSW* calculated with the very high ratio of 1.05. Although 1.05 produces much smaller values of *SSW*, the choice of discount ratio does not alter the implications of the regression estimates.

Table 1 presents values of gross and net social security wealth at six dates; the *SSW* values are converted to constant 1971 dollars by the implicit price deflator for personal consumption expenditures. The choice of discount factor is indicated by a final number in the name of the variable, for example, *SSWG*1 refers to the gross social security variable computed using a net discount ratio of 1.01. Note that by 1971 the value of *SSWG*1 was more than $2.0 trillion in constant 1971 dollars, and the value of *SSWN*1 was more than $1.1 trillion. This compares with total household assets of $3.5 trillion. The social security wealth increases household assets by 60 percent according to the gross definitions and 30 percent according to the net definition.[13]

A comparison of the *SSW* variables with real GNP (presented in col. 5 of table 1) shows that social security wealth has grown much more rapidly. This reflects a substantial aging of the population and, more important, an increase in the proportion of workers covered by social security from 58 percent of paid employees in 1940 to 89 percent in 1971.

For the regression analysis the *SSW* values are all divided by population. By construction they are already in constant 1958 dollars.

IV. Estimates of the Effect of Social Security Wealth

The estimates presented below support the conclusion that social security substantially depresses personal savings. The marginal propensity to consume social security wealth is generally statistically significant and slightly larger than the propensity to consume ordinary wealth. This difference probably reflects the fact that a large portion of ordinary wealth is held by a small fraction of households for whom bequests and the accumulation of larger fortunes are more important than saving for retirement. Although the coefficient of *SSW* depends on the choice of definition and discount rate, Section 5 will show that nearly all of the estimates imply that, in the absence of social security, personal savings would be at least 50 percent higher than they are now and probably closer to 100 percent higher.

Equation 2.1 of table 2 presents an estimate of the consumption function with gross social security wealth and a discount ratio of 1.01 (*SSWG*1).[14] The marginal propensity to consume *SSWG*1 is 0.021 with a standard error of only 0.006. The other coefficients are of the expected size: 0.650 for disposable income[15] and 0.014 for household wealth. The marginal propensity to consume gross retained earnings (0.36) is approximately equal to the value of 0.50 reported in Feldstein (1973) for an

---

[13] Although the other measures of *SSW* are lower, they are based on less acceptable assumptions about the valuing of *SSW*.

[14] Recall that the discount ratio is defined as $(1 + d)/(1 + g)$, where d is the individual's rate of discount and g is the rate of growth of real per capita income.

[15] The value of 0.650 may appear low as an estimate of the long-run marginal propensity to consume. It should be remembered, however, that the equation contains wealth and retained earnings variables. If these are omitted, the coefficient of income rises substantially.

TABLE 2

CONSUMPTION FUNCTIONS WITH SOCIAL SECURITY WEALTH

Eq.	Period*	*SSW* Definition	*YD*	YD_{-1}	*RE*	W_{-1}	*SSW*	*RU*	Const.	*SSR*	D-W Statis.
2.1	1929–71	*SSWG*1	0.530 (0.047)	0.120 (0.035)	0.356 (0.074)	0.014 (0.004)	0.021 (0.006)	. . .	228 (31)	3,618	1.82
2.2	1929–71	*SSWN*1	0.528 (0.047)	0.137 (0.034)	0.376 (0.073)	0.013 (0.004)	0.032 (0.009)	. . .	218 (27)	3,548	1.85
2.3	1929–71	*SSWG*5	0.530 (0.048)	0.136 (0.035)	0.400 (0.075)	0.008 (0.005)	0.051 (0.016)	. . .	244 (37)	3,714	1.82
2.4	1929–71	*SSWN*5	0.538 (0.049)	0.163 (0.037)	0.432 (0.079)	0.009 (0.006)	0.075 (0.027)	. . .	204 (29)	3,944	1.71
2.5	1929–71	*SSWG*1	0.675 (0.047)	0.046 (0.041)	. . .	0.009 (0.005)	0.024 (0.008)	. . .	204 (40)	6,289	1.43
2.6	1929–71	*SSWG*1	0.553 (0.050)	0.154 (0.043)	0.436 (0.096)	0.013 (0.004)	0.010 (0.011)	1.170 (0.892)	155 (64)	3,422	1.80
2.7	1929–71	*SSWG*1	0.549 (0.047)	0.149 (0.037)	0.423 (0.079)	0.012 (0.004)	0.012 (0.004)	1.020 (0.663)	169 (30)	3,430	1.82
2.8	1947–71	*SSWG*1	0.535 (0.097)	0.139 (0.097)	0.414 (0.163)	0.015 (0.009)	0.014 (0.030)	. . .	193 (159)	3,087	1.63
2.9	1947–71	*SSWN*1	0.535 (0.084)	0.119 (0.085)	0.349 (0.170)	0.014 (0.007)	0.035 (0.030)	. . .	232 (104)	2,926	1.78
2.10	1947–71	*SSWG*1	0.531 (0.098)	0.106 (0.105)	0.423 (0.164)	0.008 (0.012)	0.029 (0.035)	2.786 (3.237)	252 (175)	2,965	1.80

* The years 1941–46 are excluded from the sample. The Durbin-Watson statistic is adjusted for this gap. Standard errors are shown in parentheses.

equation that ignored the effect of social security and that was limited to the period 1929 through 1965.[16]

Using the *net* definition of social security wealth or a higher discount ratio strengthens these conclusions. Equations 2.2 through 2.4 show that alternative measures of *SSW* leave the other coefficients unchanged but increase the size of the propensity to consume social security wealth. These differences in the coefficient of *SSW* are primarily due to differences in the magnitudes of the *SSW* measures. Comparing these coefficients with the *SSW* values presented in table 1 shows that all four equations imply approximately the same effect of security wealth on consumption.

Although the retained earnings variable has a stable and statistically important coefficient, its novelty in the consumption function may raise doubts about its effect on the coefficients of the other variables. Equation 2.5 shows that excluding *RE* causes a slight increase in the coefficient of the *SSW* variable. The results are similar when a higher discount ratio is used.

Equations 2.1 through 2.5 were estimated by ordinary least squares and are therefore subject to possible simultaneous equation bias. Reestimating these equations with a consistent instrumental variable procedure confirmed the original ordinary least-squares results.[17] For example, with *SSWG*1 the coefficient of *SSW* was increased from 0.021 in equation 2.1 to 0.033 with the instrumental variable estimator. The coefficient of the retained earnings variable also rose, from 0.356 in equation 2.1 to 0.501 with the instrumental variable estimator. There was some indication of substantial simultaneous equation bias in the estimated effect of disposable income; the sum of the coefficients fell from 0.650 in equation 2.1 to 0.561 with the instrumental variable estimator.

An alternative dynamic specification of permanent income also leaves the other coefficients unchanged. More specifically, a permanent income variable was defined by the recursive relation $YP_t = (1 - \mu)YD_t + \mu YP_{t-1}$. A maximum-likelihood estimate was obtained by doing ordinary least squares conditional on values of μ between zero and one and selecting that value of μ that minimizes the sum of squared residuals. For a specification analogous to equation 2.1, the maximum-likelihood estimate of μ was 0.18, implying a rapid response that can be well represented by the 2-year distributed lag. The long-run marginal propensity to consume disposable income was 0.65, and the other coefficients were almost unchanged from the 2-year distributed lag specification of equation 2.1: 0.35 for retained earnings, 0.014 for wealth, and 0.021 for *SSWG*1.

---

[16] Since the R^2 values are extremely high for all equations ($R^2 \geq .99$), only the sum of squared residuals and the Durbin-Watson statistic are presented.

[17] The instrumental variables were government expenditure, exports, money supply, household wealth, and the social security wealth variable.

While the use of a distributed lag on past incomes is the most common way of relating consumption to "permanent" income, the rate of unemployment has often been included in the consumption function to adjust for the cyclical variation in the relation of consumption and income.[18] Introducing the unemployment rate (*RU*) in equation 2.6 lowers the coefficient of *SSW* and increases its standard error. However, the coefficient of *RU* is not statistically significant and is absolutely very small. More specifically, the coefficient implies that even a 5-percentage-point increase in *RU* would raise consumption by less than $6.00, that is, doubling the unemployment rate would raise consumption by much less than 1 percent.[19] Equation 2.7 shows that the insignificance of the unemployment coefficient is not merely due to its collinearity with the social security wealth variable. Instead of estimating separate coefficients for personal wealth and *SSW*, the two coefficients are constrained to be equal; the coefficient of *RU* is still insignificant, while the coefficient of the total wealth variable is more than three times its standard error.

The apparent collinearity of the *RU* and *SSW* variable also raises the possibility that the social security variable (which is zero before 1937) is only a reflection of the shift in consumption behavior from the prewar to the postwar period. To test this possibility, the consumption equations have been reestimated with data restricted to the postwar period (1947 through 1971). The results reject this explanation and support the original conclusion that social security substantially depresses personal savings. In equation 2.8, the coefficient of *SSWG*1 is somewhat smaller than corresponding coefficient for the entire period, while in equation 2.9 coefficient of *SSWN*1 is somewhat larger than for the entire period comparing the residual sums of squares again indicates a slight preference for the net worth measure and, therefore, the larger coefficient. Introducing the unemployment rate (eq. 2.10) substantially raises the estimated effect of *SSWG*1, while the coefficient of the unemployment variable is itself statistically insignificant. In short, although the smaller variation of *SSW* in the postwar period than in the entire sample raises the standard errors of its estimated coefficient, these coefficient estimates are similar to the values for the entire interval.

V. The Impact on Aggregate Saving

While the coefficients of table 2 must be regarded with caution, it is interesting to examine the implications of these parameter values for aggregate savings. In equation 2.1, the marginal propensity to consume disposable income is 0.650, and the marginal propensity to consume social security wealth is 0.021. In 1971, social security taxes and contributions reduced disposable income by $51 billion in 1971 prices. The corresponding reduction in personal savings is, therefore, $18 billion. The social security wealth (*SSWG*1) for 1971 was $2,029 billion, implying a reduction in saving of $43 billion. The wealth effect is, therefore, more than twice as important as the tax effect through the reduction in disposable income. The total fall in personal saving was therefore $61 billion in 1971 prices.[20] Since personal saving in 1971 was also $61 billion, the implied effect of social security is to reduce personal saving to half of what it otherwise would be.[21] Repeating these calculations for each of the other three definitions of social security wealth yields similar estimates of the reduction in personal savings, ranging from $44 billion with *SSWN*5 to $63 billion with *SSWG*5.

---

[18] See Ando and Modigliani (1963) for the rationale for using the unemployment rate and the recent level of income to represent expected future income.

[19] Real per capita consumption rose from $1,145 in 1929 to $2,393 in 1971.

The implication that social security halves the rate of personal saving is startling but not unreasonable. For middle- and low-income families, social security is a complete substitute for a substantial rate of private saving.[22] The asset-substitution effect is, therefore, likely to be very significant. Although social security induced retirements that would otherwise have been postponed, a substantial fraction of older men were already retiring before the introduction of social security. In 1930, 46 percent of men over 65 were retired, while in 1971 this was 75 percent. For the original 46 percent, social security has only an asset-substitution effect. Only for the remaining 25 percent does social security have a stimulating "retirement effect" as well as a depressing asset-substitution effect.[23] The estimates imply that any additional savings that result from induced retirement among this group have been outweighed by the asset-substitution effect among all social security beneficiaries.

Although the coefficient of social security wealth has been treated as a constant in estimating the consumption function, it is more appropriate to regard the estimate as an average of the different propensities to consume that prevailed during the sample period. Moreover, it seems likely that the marginal propensity to consume increases in *SSW* will be higher in the future than it has been in the past. The coefficient of *SSW* reflects both the asset-substitution effect and the inducement-to-retire effect. The relative importance of these two countervailing effects will change as the social security program matures. Until recently, social security has been increasing the frequency of retirement among men over age 65. These induced retirements led to additional saving and, therefore, partly offset the asset-substitution effect that gives the *SSW* coefficient its positive sign. Since 75 percent of men and 91 percent of women over 65 were retired in 1971, there is now much less scope for a proportional increase in retirement, and the asset-substitution effect can be expected to dominate even more strongly in the future.[24]

---

[20] This slightly overstates the effect on saving because the social security taxes and contributions include funds for Medicare and disability insurance. The calculations implicitly assume that the parameters of the consumption function would be unchanged if consumer expenditure were replaced by the sum of consumer expenditure, interest paid by consumers, and personal transfers to foreigners; the two additional items were only 3 percent of 1971 consumer expenditure.

[21] Personal saving in 1971 was 9.2 percent of disposable income and thus substantially higher than other years in the postwar period. For the postwar sample period, the average was 6.4 percent. The choice of 1971, therefore, substantially understates the relative impact of social security. Note, however, that the assumption that all social security benefits are consumed may overstate the effect of the program on saving. If instead the propensity to consume these benefits is assumed to be the same as for other forms of disposable income, only the wealth effect would remain. This implies that saving would be reduced by $43 billion to 59 percent of what it would otherwise be.

[22] For an individual with earnings below the covered maximum ($9,000 in 1972), social security can replace personal savings of at least 10 percent of disposable income.

[23] Note that the 46 percent and 25 percent refer to man-years and not to individuals.

VI. Some General Implications

The model and empirical estimates presented in this paper have wide-ranging and important implications. Although these cannot be explored here in any detail, the following section will provide some brief remarks about three different issues.

The Theory of Consumption

The analysis lends support to the general idea of life-cycle saving by indicating the importance of extending the traditional model by making the age of retirement an economic decision. The substantial estimated effect of social security wealth on consumption is consistent with this life-cycle hypothesis. By taking into account the effect on saving of induced retirement, the generalized life-cycle model can also explain previous survey evidence on the relation between pensions and household saving without recourse to Cagan's "recognition effect" or Katona's "goal gradient" hypothesis.

Previous tests of the life-cycle model should be reconsidered within this more general framework and with explicit recognition of the role of social security.[25] A measure of social security wealth should also be included in future tests of other hypotheses about consumption behavior.

---

[24] Munnell (1973) independently recognized the important effect of changing retirement behavior on aggregate savings. Her study examines the various forces that have altered retirement rates since 1900 and provides an explicit estimate of the positive effect of increasing retirement on personal saving. By using national balance-sheet data to construct savings series, she is able to separate savings in the assets that are most important for the families with low and moderate incomes for whom social security is a primary source of retirement income. Her analysis using the *SSW* series of the current paper provides strong support for the conclusion of this section.

One of the most firmly established facts about consumer behavior is the long-run constancy of the aggregate savings rate (Kuznets 1946; Denison 1958; David and Scadding 1972). Much of modern consumption theory arose in the attempt to reconcile this fact with other observations about aggregate and individual consumption. The substantial depressing effect of social security on the postwar savings rate raises a serious question about the *behavioral* significance of the observed constancy of the savings rate. The evidence presented in this paper seems more consistent with the Keynesian view that the aggregate rate of saving would increase as income rose if there were no offsetting government policies.

Capital Accumulation and National Income

The evidence that the social security program approximately halves the personal savings rate implies that it substantially reduces the stock of capital and the level of national income. During the 1960s, personal saving accounted for 60 percent of total private saving. By halving personal saving, social security reduced total private saving by 38 percent. In the long run, this decrease in the rate of private saving would also decrease the private capital stock by 38 percent.[26]

These estimates are consistent with the magnitude of the social security wealth derived in Section 3. The 1971 value of gross social security wealth with the most plausible value of the relative discount ratio ($SSWG1$) was $2,029 billion in 1971 prices. Total household assets (W) in that year were $3,474 billion in 1971 prices. The actual level of assets was, therefore, 37 percent less than if all of the social security wealth had instead been saved privately. This is remarkably close to the predicted long-run effect of 38 percent, in spite of conceptual differences between household assets and real capital stock.

A 38 percent decrease in the capital stock implies a substantial reduction in GNP. If this asset substitution had not occurred, the long-run capital stock would be some 60 percent higher.[27] A rough approximation of the effect that this greater capital stock would have on national income

[25] The possibility that the generalized life-cycle model is appropriate for most of the population but not for the very wealthy should be examined. See Tobin (1967) for an interesting simulation approach to testing the consistency of the life-cycle hypothesis with actual savings, and Mayer (1972) for a survey of previous econometric estimates.

[26] Since the social security system is not "funded" but operates on a pay-as-you-go basis, the reduction in private saving is not offset by any increase in public saving. Note that if the tax-transfer process does not reduce saving so that saving falls only because of the wealth effect of the social security program, the decrease in the rate of private saving and therefore in the private capital stock is 29 percent.

[27] In providing a "fiat" asset as a substitute for real capital formation, social security is similar to absorbing saving by a growing money supply (see Tobin 1965).

can be obtained by assuming a Cobb-Douglas technology and a capital coefficient of 0.3. With this assumption, GNP is multiplied by $(1.6)^{0.3}$, that is, income would rise by 15 percent. A more conservative assumption of a 40 percent increase in the capital stock implies that GNP would rise by 11 percent. For 1972, this implies that GNP would be increased by more than $127 billion.

The lower level of GNP reflects the pay-as-you-go nature of our social security system. Because social security contributions are used to pay concurrent benefits, the capital stock is smaller and income is less.[28] This result should be contrasted with Samuelson's (1958) important and much misinterpreted analysis of the effect of social security in a "consumption loan" model. Samuelson considered an economy in which real capital accumulation was physically impossible. In his analysis, all commodities were like "chocolates" that melted and could not be stored through time. Private saving required finding a borrower who wished to consume more than his income and could repay at a later time. In this situation, life-cycle saving cannot be carried out efficiently. Samuelson showed that a pay-as-you-go social security system could lead to an unambiguous improvement in welfare; each generation would prefer the social security system to the alternative equilibrium. In this analysis, the assumption that there is no store of value or method of real accumulation is crucial. As Samuelson notes, simply introducing a stock of fiat mon provide a store of value makes it unnecessary to create a social sec system in order to achieve intertemporal efficiency. Allowing for accumulation of productive capital goods makes the analysis more c plex and, as shown here, implies that a pay-as-you-go social sec system reduces aggregate saving and lowers the level of real income.

The Distribution of Income

Analyses of the distributional impact of social security have focused on two questions. First, how do the taxes nominally paid by persons at different income levels compare with the benefits that are eventually received?[30] Second, how does the payroll tax affect gross factor prices, that is, what is the incidence of the tax? This paper indicates that the benefits side of the social security program also has important effects on the distribution of income. Two separate issues can be distinguished: the changes in factor prices, and the effects on the income of the aged.[31]

---

[28] Note that this lower level of income does not measure lost welfare because some additional consumption was obtained at an earlier date. An analysis of the welfare loss would require a cardinal social utility function capable of intergenerational calculations of time preference.

[29] As Aaron (1966) has noted, a pay-as-you-go social security system can lead to an unambiguous welfare increase in an economy that would otherwise save too much, that is, in the unlikely case of an economy that is so capital intensive that the rate of return is less than the rate of growth of real output.

[30] The system has been shown to pay a rather high implicit rate of interest on these taxes, with relatively higher yields going to those at lower income levels (see Brittain 1972). The redistributive impact is generally understated by ignoring the fact that higher-income individuals are much less likely to retire at age 65.

The substantial reduction in the capital stock that results from the current social security program decreases wage rates and increases the rate of profit. To obtain an order of magnitude for these effects, I will again assume a Cobb-Douglas technology with a capital elasticity of 0.3. With a fixed labor supply, the wage rate is proportional to $K^{0.3}$ (where K is the capital stock), and the rate of interest is proportional to $K^{-0.7}$. If, in the absence of the current social security system, the capital stock were 60 percent greater, the wage rate would rise by 15 percent and the rate of interest would fall by 28 percent. Although the Cobb-Douglas technology implies that factor shares would be unchanged, there would be a substantial redistribution of capital income to those who, because of the social security system, now save very little.

The induced retirement implies that the labor supply is also decreased by the social security system. Less labor raises the wage rate and lowers the rate of return.[32] This partly offsets the distributional effect of the change in the capital stock. But even if all of the increase in retirements since 1930 were attributed to social security, the reduction in the labor supply would be quite small in comparison to the reduction in the capital stock. More specifically, if the labor force participation rates of those over 65 were at the 1930 values, the labor force in 1970 would be increased by less than 3 percent.[33] The net effect of the changes in capital and labor has clearly been to raise the rate of profit and lower the wage rate.

Although the social security program was intended primarily as a system of income maintenance for the aged, it may now have the effect of reducing the levels of income and consumption for many of those over 65. Those who are induced by the combination of available benefits and the "earnings test" to retire earlier than they would otherwise have done generally receive substantially less in social security benefits than they would have earned if they had continued to work. As the analysis of Section 1 showed, even with optimal savings during the preretirement years, the level of consumption is likely to fall at retirement. The income-maintenance goal of the social security program is thus partly vitiated by imposing an earnings test as a condition for receiving benefits.[34]

---

[31] By replacing individual saving, social security also has a substantial effect on the distribution of wealth. If each household's social security wealth were added to its other assets, the distribution of total wealth would be much more equal than the distribution of wealth as traditionally defined.

[32] A reduction in labor supply through earlier retirement has a different long-run effect on factor prices than a proportional reduction in the quantity of labor supplied each year with no change in retirement. In particular, these two changes in labor supply have different effects on savings behavior (see Feldstein 1972).

[33] Future increases in the level of social security benefits will have relatively little effect on the labor supply since 74 percent of men over 65 are already retired.

VII. Conclusion

It is best to conclude this paper with a caveat. The substantial estimated effect of social security on personal savings is based on aggregate time-series data for a single country. The results are consistent with the generalized life-cycle hypothesis but require further analysis with different sets of data. Work is currently in progress to examine household data for the United States and aggregate data for a cross-section of countries. Additional studies for individual countries, using both aggregate time series and household surveys, could be of great value in testing the current conclusions.

References

Aaron, Henry. "The Social Insurance Paradox." *Canadian J. Econ. and Polit. Sci.* 33 (August 1966): 371–74.

———. "Social Security: International Comparisons." In *Studies in the Economics of Income Maintenance,* edited by Otto Eckstein. Washington: Brookings Inst., 1967.

———. "International Comparisons," appendix D. In *Social Security: Perspect*
Reform, by J. A. Pechman, H. J. Aaron, and M. K. Taussig. Wash
Brookings Inst., 1968.

Ando, Albert, and Modigliani, F. "The 'Life Cycle' Hypothesis of S
Aggregate Implications and Tests." *A.E.R.* 53 (March 1963): 55–84.

Brittain, John A. *Corporate Dividend Policy.* Washington: Brookings Inst., 1

———. *The Payroll Tax for Social Security.* Washington: Brookings Inst., 19

Cagan, Phillip. *The Effect of Pension Plans on Aggregate Savings.* New York: Nat. Bur. Econ. Res., 1965.

David, P., and Scadding, J. L. "What You Always Wanted to Know about 'Denison's Law' But Were Afraid to Ask." Memorandum no. 129, Stanford Center for Research, 1972.

Denison, Edward F. "A Note in Private Saving." *Rev. Econ. and Statis.* 40 (August 1958): 261–67.

Feldstein, M. S. "Corporate Taxation and Dividend Behavior." *Rev. Econ. Studies* 37 (January 1970): 57–72.

———. "Tax Incidence in a Growing Economy with Variable Factor Supply." Harvard Institute of Economic Research Discussion Paper no. 263, 1972. Forthcoming in *Q.J.E.*

———. "Tax Incentives, Corporate Saving, and Capital Accumulation in the United States." *J. Public Econ.* 2 (April 1973): 159–71.

———. "Social Security and Private Savings: International Evidence in an Extended Life Cycle Model." Harvard Institute of Economic Research Discussion Paper, 1974. Forthcoming in *The Economics of Public Services: Proceedings*

---

[34] There is also an important loss of economic welfare due to induced retirement. The earnings test and the social security benefits imply a very high marginal rate of tax on earnings after age 65.

of an International Economics Association Conference, edited by Martin Feldstein and Robert Inman.

Feldstein, M. S., and Fane, George. "Taxes, Corporate Dividend Policy and Personal Savings: The British Postwar Experience." *Rev. Econ. and Statis.* 55 (November 1973): 399–411.

Fisher, I. *The Theory of Interest*. New York: Macmillan, 1930.

Friedman, Milton. *A Theory of the Consumption Function*. Princeton, N.J.: Princeton Univ. Press, 1957.

Goldsmith, R. W. *A Study of Saving in the United States*. Vols. 1–3. Princeton, N.J.: Princeton Univ. Press, 1956.

Harris, Seymour. *Economics of Social Security*. New York: McGraw-Hill, 1941.

Harrod, R. F. *Towards a Dynamic Economics*. London: Macmillan, 1948.

Katona, George. *Private Pensions and Individual Savings*. Ann Arbor: Survey Research Center, Institute for Social Research, Univ. Michigan, 1964.

Kuznets, Simon. *National Product since 1869*. New York: Nat. Bur. Econ. Res., 1946.

Mayer, Thomas. *Permanent Income, Wealth and Consumption*. Berkeley: Univ. California Press, 1972.

Modigliani, Franco. "The Life Cycle Hypothesis of Saving and Intercountry Differences in the Saving Ratio." In *Induction, Growth, and Trade: Essays in Honour of Sir Roy Harrod*, edited by W. A. Eltis, M. FG. Scott, and J. N. Wolfe. Oxford: Clarendon, 1970.

———. "Monetary Policy and Consumption: Linkages via Interest Rate and Wealth Effects in the FMP Model." In *Consumer Spending and Monetary Policy*. Monetary Conference Series no. 5. Boston: Fed. Res. Bank Boston, 1971.

Modigliani, F., and Brumberg, R. "Utility Analysis and the Consumption Function: An Interpretation of Cross Section Data." In *Post Keynesian Economics*, edited by K. Kurihara. New Brunswick, N.J.: Rutgers Univ. Press, 1954.

Munnell, Alicia. *The Effect of Social Security on Personal Savings*. Ph.D. dissertation, Harvard Univ., 1973.

Pechman, Joseph A.; Aaron, Henry; and Taussig, Michael K. *Social Security: Perspectives for Reform*. Washington: Brookings Inst., 1968.

Samuelson, Paul. "An Exact Consumption Loan Model of Interest with or without the Social Contrivance of Money." *J.P.E.* 66 (December 1958): 467–82.

Tax Foundation. *Economic Aspects of the Social Security Tax*. New York: Tax Foundation, Inc., 1967.

Tobin, James. "Money and Economic Growth." *Econometrica* 33 (October 1965): 671–84.

———. "Life Cycle Saving and Balanced Growth." In *Ten Economic Studies in the Tradition of Irving Fisher*, edited by W. Fellner et al. New York: Wiley, 1967.

U.S. Department of Health, Education and Welfare. *Life Tables: 1959–61*. PHS Publication no. 1252. Vol. 1, no. 1. Washington: Government Printing Office, 1968.

Public Goods, Perfect Competition, and Underproduction

William H. Oakland

Ohio State University

Recently, the classical position that free markets will underprovide public goods has come under attack. Critics have argued that, if exclusion is costless, the proposition is false. Underlying this argument, however, is the assumption that preference maps of consumers are known. In this paper, this assumption is relaxed. It is shown that a system of atomistic competition will lead to suboptimal levels of public goods, and that those public goods which are produced will be inefficiently allocated. Furthermore, owing to different intensities of use, units of public goods will command different prices. However, for any particular unit of public good, price will be the same for all individuals.

I. Introduction

The question of how a market economy, left to its own devices, allocate public goods has received considerable attention in recent years. With a few notable exceptions, the literature has focused upon that subset of public goods for which exclusion is not possible, or, if possible, prohibitively costly. This includes the case of pure public goods, such as defense, and impure public goods, as represented by production and consumption externalities. The consensus of such studies is that private markets will systematically underprovide for collective goods.[1]

Recently several investigators (Demsetz 1970; Thompson 1968) have explored the consequences of market provision of public goods for which exclusion is costless. It is surprising that each author arrives at the conclusion that competitive markets will not underprovide public goods. Indeed, Thompson (1968) arrives at the conclusion that overprovision will emerge. Both authors also conclude that competitive equilibrium will

---

An earlier version of this paper was read to the March 1973 meetings of the Public Choice Society. I am grateful to Wallace Oates and James Buchanan for useful suggestions.

[1] See, for example, Mishan (1971), Musgrave (1959), and Samuelson (1954). For an exception to this statement, see Buchanan and Kafoglis (1963).

[*Journal of Political Economy*, 1974, vol. 82, no. 5]
© 1974 by The University of Chicago. All rights reserved.

be characterized by price differentials among individuals for the public good.

It is not the purpose of this paper to offer a critical review of the merits of the aforementioned studies.[2] Instead, our aim is to relax one of the key assumptions made by both authors—that preferences of individual consumers are known to sellers of the public good.[3] Since such an assumption is not generally made with respect to private goods and has, to date, been the root of difficulties confronting policy makers with respect to public goods, it seems interesting and important to relax it. Such a procedure, we feel, will lead to a more adequate framework for dealing with the question of the competitive supply of public goods for which exclusion is not prohibitively costly.

The results of our analysis are similar in several important respects to those already published. Needless to say, they also differ in important respects. As in most previous studies, but in direct contradiction to at least two—Owen (1969) and Ekelund and Hulett (1973)—competition is shown to lead to a determinant, nonzero output of the public good. Furthermore, price differentials exist in equilibrium. However, unlike the case in existing studies, prices here differ among units of the public good and not among individuals. Competitive equilibrium is shown to depart from Pareto optimality for two reasons: (1) certain individuals may be excluded from consuming a portion of the output of public goods; (2) the level of production of public goods is too small. Thus, we are able to demonstrate that, in the absence of firms' knowledge concerning the preferences of specific individuals, the orthodox underprovision doctrine carries over to the case where exclusion is costless.

II. The Model

Consider an economy composed of N individuals and two goods, A and B. Let good A represent a private good and good B a public good. Good B has the property that each of the N individuals in the economy can consume its total production,[4] although any individual can be costlessly excluded. We shall assume that production in the economy is characterized by constant costs—that is, the production-possibility curve is linear.[5]

---

[2] For a critique of Thompson's paper, see Ganguly (1969), Owen (1969), and Rodgers (1969); see also Thompson's reply (1969).

[3] Thompson (1968) assumes that the preference map of each consumer is known to both sellers and other consumers. Demsetz (1970), on the other hand, makes the weaker assumption that individual demand schedules for the public good are known to sellers.

[4] For concreteness, we can think of the public good as the production of television tapes, among which individuals are indifferent; transmission costs are taken to be zero, and scrambling devices can be costlessly installed on receivers. This example is given by Demsetz (1970).

[5] This assumption is made for expositional purposes only. We need only require that the appropriate convexity conditions hold.

We shall take the price of good A to be fixed at one so that all values can be expressible in terms of units of A. Entry into the production and sale of both goods is assumed to be free, and we assume that the market for A is perfectly competitive in the usual sense of the term. Finally, each individual, i, is assumed to have a strictly quasi-concave utility function, u^i, defined over the two goods, that is, $u^i = u^i(A^i, B^i)$, where A^i and B^i correspond to the ith individual's consumption of A and B, respectively.

To begin with, we shall assume that, in the market for B, each individual behaves parametrically in terms of price. Hence, we can define the inverse-demand function

$$p^i = f^i(B^i) \qquad (i = 1,\dots, N) \tag{1}$$

for each individual. To arrive at (1), each individual maximizes his utility subject to his budget restraint. Since the consumption of B is nonrival, we obtain the market demand, R, by summing the individual demand functions vertically:

$$R = \sum_{i=1}^{N} P_i = \sum_{i=1}^{N} f^i(B). \tag{2}$$

Equation (2) corresponds to the marginal valuation of the community for any quantity B of the public good.

Turning to the supply side, we assume that firms have no knowledge of the preferences of specific individuals. Since marginal cost of extendin consumption to another individual is zero, each firm is assumed to in every individual to consume its entire output; firms will enter the ind as long as unit price as given by (2) exceeds the constant cost of produc which we represent by c. Hence, the supply function for the econom

$$R = c.$$

Combining (2) and (3), we obtain the equilibrium condition for th economy:

$$\sum_{i=1}^{N} f^i(B) = c. \tag{4}$$

Together, equations (1) and (4) define the equilibrium values of B and the p^i. It can be easily shown that this solution is Pareto-optimal.[6]

One of the interesting features of the model just described is that in equilibrium each individual pays a different price for the public good and consumes the same quantity. This is precisely the reason why the solution described by (1) and (4) is likely not to be achieved. As Samuelson (1969) so aptly phrased it, the above only represents a "pseudo-equilibrium."

It will be recalled that the solution defined by (1) and (4) was based upon the premise that consumers behaved parametrically with respect to price and that all firms attempt to sell their entire output to all individuals. It is easy to see, however, that if the number of individuals is large, such parametric price behavior of consumers is not based on rationality. Inspection of (1) and (4) reveals that it is in the interest of each individual to reduce his price offer. Supply decisions are based upon $\sum P^i$, of which any P^i can be assumed to constitute a negligible fraction. Thus, the rational individual will behave parametrically with respect to quantity and not with respect to price. In other words, each individual thinks of himself as a monopsonist who is faced with a perfectly inelastic supply schedule. If preferences are not known, rational consumers will find it in their own interest to misrepresent their preferences by making price offers substantially below those corresponding to (1).[7] Indeed, in the limit, each consumer will express no demand whatever for the collective good, irrespective of the price. Each individual is caught in the well-known "prisoner's dilemma." The result will be little or no private production of the public good, which is exactly what orthodox underprovision theory predicts.[8]

---

[6] Readers will recognize the similarity between the present model and the Lindahl theory of voluntary exchange; see Johansen (1965, chap. 6).

It was perhaps in response to this dilemma that Demsetz (1970) and Thompson (1968) introduce the assumption that preferences are known. For it might seem that the individual buyer is able to exploit his monopsonistic position only by misstating his preferences. As long as perfect competition is assumed on the supply side, however, it is not clear that the knowledge of preferences will have any impact on the outcome. It is also not clear that a model built upon such a premise can be described as "competitive." Whatever the case, however, the assumption is not necessary in order to develop a model that produces a nonzero equilibrium output of the public good. This will be shown in what follows.

III. An Alternative Approach

Thus far, our discussion of the competitive production of a public good has been based upon the premise that a particular unit of public good can command a different price for different individuals. The motivation for this assumption of a public good to an additional person is zero. Thus, profit-maximizing firms, taking unit prices as given, will attempt to induce each consumer to purchase their entire output. Since individual tastes and income vary, the latter can be accomplished only by charging different persons a different price. In other words, in order to clear the *N* markets for the public good, *N* independent prices are generally required. Thus, price differentials among consumers are a requirement for competitive equilibrium.

---

[7] See, however, the recent experiments of Bohm (1972), which indicated that individuals may voluntarily reveal their preferences for public goods even if cost shares are dependent upon their response. However, these experiments were carried out in a small-number setting, and thus may be of limited relevance to the problem at hand.

[8] Similar arguments concerning the implausibility of the model have been set forth by, among others, Buchanan (1968), Musgrave (1959), Olson (1965), and Samuelson (1954, 1958, 1969). It has also been described by Owen (1969) in the present context. Ekelund and Hulett (1973) arrive at a different conclusion regarding the outcome. They argue that a process of price cutting would lead to a single supplier of the public good. Freedom of entry would limit profits of the remaining firm to normal levels. However, I find their argument unconvincing. It is based upon the allegation that a firm whose revenues and costs initially satisfy eq. (4) can profit by luring customers away from other firms. Since firms are assumed to be selling their entire output to all individuals, this clearly cannot be the case.

On the other hand, the requirement that each person purchase total industry output leads to the situation where each person is a monopsonist in the market for public goods. From the point of view of the individual, this monopsony power is so strong that it leads him to attempt to consume public goods free of charge. Such behavior would appear to preclude the production of public goods in a competitive framework. Thus, from the point of view of establishing a viable competitive equilibrium for public goods, little is gained by assuming price discrimination among individuals. In the analysis to follow, we shall attempt to develop a competitive equilibrium under the assumption that the price for the services provided by a particular unit of the public good is identical for all persons. In this case, a consumer cannot gain by misrepresenting his true preferenc because from his point of view such activities will have a negligible eff upon the price he pays. Price uniformity across individuals for a given un will clearly be the case if resale of consumption rights is permitted. It wi also prevail because we assume that firms are unable to distingui individual demand curves. The only knowledge firms have concernin individual tastes is whether or not a particular individual chooses to purchase the services of their own output. This is clearly an extreme assumption since, for many services, costs of determining something about a particular consumer's demand function may not be prohibitive. For example, occupation or housing characteristics may be correlated with demand elasticity. Nevertheless, we feel the polar case of consumer indistinguishability is of considerable interest. It is the assumption most frequently employed in the analysis of private goods. Furthermore, the preference-revelation problem is generally regarded as the most serious obstacle to the efficient provision of public goods. By making an extremely weak assumption about the knowledge of consumer preferences, we are able to isolate the role of exclusion in providing such information. Finally, existing studies have been based upon an equally strong, but opposite, assumption about knowledge of consumer tastes. Our results can thus be used to define the range of possible outcomes. It is likely that intermediate cases will carry the flavor of both extremes.

Despite the fact that the price of a given unit of public good will be the same for all individuals, it does not follow that the prices of different units will be identical.[9] Indeed, it will be shown that, in general, such price differentials tend to emerge from the competitive process. In the following discussion, it will be convenient to think of each firm as producing a single unit of collective good. Given the assumption of constant costs of production, this assumption is innocuous.

Consider, first, the supply side of this economy. It is clear that a particular unit of public good will be produced if and only if revenues cover costs; that is,

$$P \geq c/n, \tag{5}$$

where n denotes the number of individuals consuming the unit ($0 < n \leq N$), and P represents the price paid by each individual consuming the good. Because of our assumption of free entry, it must be the case that the equality in (5) prevails. Thus, we have

$$P = c/n. \qquad (n = 1, 2, 3, \ldots, N). \tag{6}$$

This relation states that all units achieving an intensity of use n must be sold at the same unit price P_n, and the latter must just cover average costs c/n. Thus, to the extent that units of collective goods are used at different intensity levels, their prices will differ. This case is in direct contrast with earlier models, where each firm is assumed to receive the same unit price and all individuals necessarily consume all units.

Turning to the demand side, it is clear that individuals will be subjectively indifferent between units of public goods having differing intensities of use. Therefore, they will purchase those units of B that are cheapest. The lowest price they will face will be c/n and the highest price they may face is c. In general, they may face as many as N different prices with different quantities of public goods available at each price. Let the function $x_n = h(P_n)$ represent the number of units available at price P_n. The function h is a step function since, in equilibrium, the Ps can only assume the discrete values defined by (6). The problem facing each consumer can be described as follows:

$$\begin{aligned} &\max u^i(A^i, B^i), \\ &\text{subject to: } \sum_{j=1}^{N} v_j^i = B^i, \\ &\qquad \sum_{j=1}^{N} P_j v_j^i + A^i = \bar{A}^i, \\ &\qquad 0 \leq v_j^i \leq x_j = h(P_j), \qquad j = 1, \ldots, N \\ &\qquad A^i \geq 0, \end{aligned}$$

---

[9] Throughout the paper, the term price refers to that amount paid by a consumer to enjoy the services of a particular unit of public good. This should be distinguished from that amount received by the firm per unit of public good, $\sum p^i$.

where v^i_j represents the quantity collective good available at price P_j which is consumed by the ith individual. This is a well-defined nonlinear programming problem. The Kuhn-Tucker conditions generate the following relations:

$$P_j = MRS^i, \qquad 0 < v^i_j < x_j \tag{7}$$

$$P_j \geq MRS^i, \qquad v^i_j = 0 \tag{8}$$

$$P_j \leq MRS^i, \qquad v^i_j = x_j$$

$$MRS^i = \frac{\delta u^i}{\delta B^i} \bigg/ \frac{\delta u^i}{\delta A^i}, \tag{9}$$

where MRS^i is the marginal rate of substitution for the ith individual. Inspection of these conditions quickly reveals that (7) can hold for no more than one P_j, and if it holds, the strict inequality must hold in (8) and (9). In other words, the consumer will consume the maximum available at a given price before consuming units at a higher price. Combining (7), (8), and (9) yields the condition

$$P_j \leq MRS^i \leq P_{j-1}, \tag{10}$$

where P_j is the highest price at which the individual trades. Together with the consumer's budget constraint, (10) defines the ith person' demand $\hat{B}^i$.

Consider the ordered vector $\hat{B} = (\hat{B}^1, \hat{B}^2, \ldots, \hat{B}^N)$, where the superscripts i are chosen such that the $\hat{B}^i$ are arranged in descending order The conditions for equilibrium are:

$$\begin{aligned} &\hat{B}^N = x_N, \\ &\hat{B}^{N-1} - \hat{B}^N = x_{N-1}, \\ &\quad\vdots \\ &\hat{B}^{N-k-1} - \hat{B}^{N-k} = x_{N-k-1}, \\ &\quad\vdots \\ &\hat{B}^1 - \hat{B}^2 = x_1. \end{aligned} \tag{11}$$

The task then, is to find a step function $h(P_n)$ which will satisfy (11). A sketch of the proof that such a function exists is as follows: Set x_N equal to $\hat{B}^i(\text{min})$ which is arrived at by solving the consumer's problem when the price is fixed at c/N; set x_{N-1} equal to $\hat{B}^i(\text{min})(i \neq N)$ when the price of the first x_N units is c/N and the price of all others is $c/N - 1$; etc.

Several characteristics of this solution are worth noting. First, the number of different prices for which nonzero quantities are traded is equal to the number of distinct consumer demands $(\hat{B}^i)$. If all individuals are identical in tastes and resources, each unit of public good will be sold at $P_N = c/N$.[10] If, on the other hand, each individual's tastes and real incomes are unique, trade may occur at as many as N different prices.[11] In general, therefore, some form of price discrimination will emerge from the process of competition. Such price discrimination takes the form of different prices for different units of collective good. Units of public good consumed by relatively few individuals will command a relatively higher price. Despite the fact that resale of consumption rights is permitted, arbitrage among consumers cannot eliminate these price differentials. This owes to the fact that lower-priced units are already consumed by those who purchase the higher-priced units and a given unit cannot be consumed more than once by a particular consumer.

Second, units sold at a price exceeding c/N will be characterized by "excess capacity"; that is, certain additional consumers could be accommodated at zero cost. There is, however, no incentive for the firm to lower its price to attract these consumers. For in order to do so, it would suffer a revenue loss on its existing clientele of an amount exactly equal to the revenue received from the new customers. Thus, the solution described by (6) and (11) is a stable equilibrium.

Third is the unusual industry structure of firms in the public goods sector. Firms are neither price takers nor price setters in the usual sense. Instead, they face a schedule of prices and associated sales that they take as given. While they may choose which point on the schedule to operate, they are indifferent among alternative strategies. Freedom of entry insures that the actions taken by an individual firm have no influence on the price-sales schedule they face. It is such parametric behavior that gives the present model its strong competitive flavor.

Fourth is the key role played by the assumption that firms do not know the preferences of particular individuals. Recall, a firm selling to n customers, $n < N$, would find it profitable to attract new customers if it could price discriminate between its existing and new clientele. Clearly, arbitrage would rule out this possibility. However, even if the firm were able to prevent arbitrage, it is unable to distinguish among its customers with respect to demand price. Hence, if it were to give one individual a price break, the same treatment would have to be extended to all of its customers—thereby negating the benefit of the additional customer. It is important to note that the model does not imply that consumer preferences are not revealed in the marketplace. Indeed, such revelation is a necessary condition for the existence of private markets. What is assumed in the present paper is that preference orderings of specific individuals are not known by firms—an assumption certainly not required by the standard competitive model.

---

[10] In this case our conclusions are the same as Demsetz's (1970).

[11] That each individual differs in tastes and real income is a necessary, but not sufficient, condition for the emergence of N different prices.

The nature of the equilibrium and the roles played by key assumptions can be clarified by appeal to a particular example, television tapes. Let us assume that television tapes are perfectly substitutable for one another and that no consumer wishes to view the same tape twice. Furthermore, tapes are assumed to be produced at constant cost. According to the solution described in (11), then, the price for viewing tapes will vary among the different tapes, but will be the same for all viewers of a given tape. Furthermore, utilization rates will also vary among tapes—the more expensive being the less fully utilized, and vice versa.

Since entry is free and arbitrage possible, how can the price differences among tapes be sustained? Will not competition and the fact that marginal cost is zero for nonfully utilized tapes force price to its lowest possible level, c/N? The answer to these questions is no, as we shall demonstrate.

Freedom of entry means that firms are willing to supply any quantity of tapes as long as they can cover their production costs—c. This means that a firm is indifferent to supplying a tape to n users ($1 \leq n \leq N$) or $n + k$ users ($1 \leq n + k \leq N$) if the prices it can receive are c/n and $c/n + k$, respectively. If any tape, carrying price c/n, is sold to more than n customers, the resulting profit will attract competitors, forcing the price down to a level consistent with zero profit. In other words, the elasticity of supply for tapes used by n different people is infinite at c/n What remains, then, is to show that different tapes will have different numbers of subscribers. Consider first the price c/N. This is the lowest achievable price because the cost of the tape is spread over the largest possible number of consumers. At this price, consumers, unless identical in tastes and income, will want to consume at different levels. Let x_0 correspond to the smallest individual demand at price c/N. Then freedom of entry will assure that x_0 tapes will be produced and that each tape will be sold to all households.

What about those people who wanted to view more than x_0 tapes at price c/N? Suppose we consider the person with the second-smallest demand at price c/N, say x_1 tapes. Under what conditions will the $(x_1 - x_0)$ tapes be produced? Clearly, we must require that the costs be covered. If there are $(N - 1)$ people who have demands at least as great as x_1, at price c/N, the lowest possible price for the $(x_1 - x_0)$ goods is $c/(N - 1)$. Higher prices will be beaten down by free entry; lower prices will be unprofitable.

So far, we have x_0 tapes being sold to all consumers at price c/N and $(x_1 - x_0)$ being sold to all but one consumer at price $c/(N - 1)$. Can such a price differential be sustained? Consider the supply side. Firms selling to all individuals make no more of a profit than those selling to $(N - 1)$ individuals. If a firm that presently sells to all individuals raises its price above c/N, it will lose his entire clientele to a new entrant. By similar reasoning, a firm selling to $(N - 1)$ people cannot increase its price above $c/(N - 1)$, nor will it lower its price to below $c/(N - 1)$. To offset the reduction in price, such a firm would have to attract an additional customer. By definition, however, the necessary price would be c/N. If the firm could charge the lower price to only the one individual, it would find this profitable. However, arbitrage among the consumers of a given tape rules out this possibility. More important, however, is the fact that each of the original $(N - 1)$ customers could demand the same treatment, and freedom of entry would ensure these demands would be met. Hence, to attract an additional consumer, firms selling to only $(N - 1)$ individuals would have to lower their price for all individuals to c/N—negating the benefit of the additional customer.

Finally, we can ask whether arbitrage among consumers can eliminate any price differentials between units. Since consumers will purchase the services of lower-price tapes before higher-priced units, an individual who watches higher-priced tapes is consuming all lower-priced tapes. Hence, individuals who buy at low prices cannot sell to those buying the higher-priced tapes because the latter are already viewing the low-priced tapes. Thus, arbitrage among tapes of different prices is not possible.

IV. Optimality Properties

It is obvious that, in general, the solution given by (11) does not satisfy the requirements of a Pareto-optimum allocation. The mere fact that some individuals are excluded from consuming part of the output of collective goods is sufficient to preclude Pareto optimality. The extent of the deadweight loss will depend upon the degree of diversity of consumer tastes and resources. Only in the limiting case of identical individuals does this source of inefficiency (as well as any other) disappear. That the use of average-cost prices to allocate collective goods will result in inefficiency is a well-known proposition and dates as far back as Dupuit's analysis of the uncongested bridge.

What can be said concerning the optimality of the level of public good production implied by (6) and (11)? In order to answer this question, it might appear that we would have to define the optimal level of collective goods output and compare it with the level which satisfies (6) and (11). In general, however, the Pareto-optimal level of public goods production is not unique but depends upon the initial distribution of income and the way in which the costs of the collective good are allocated. It is quite possible, therefore, that our answer to the question posed above will be ambiguous. This is not the sense, however, in which the terms under- or overproduction have been used by economists. Consider, for example, an often-used explanation of the inefficiency involved in the case of monopoly. Suppose all but one market satisfy the conditions of Pareto optimality; in this case, marginal cost corresponds to true opportunity costs. Since price exceeds marginal cost and since the former is an indicator of marginal personal benefit, it is usually alleged that monopoly results in the underproduction of the good in question (and by implication the overproduction of the other goods, taken together). The reasoning is that the consumer can be made better off and the firm no worse off if an additional unit of good is produced and made available to the consumer at a price equal to marginal cost. Absent from this argument is any explanation of how such a change is brought about. Furthermore, it ignores any income effects which result from the increase in the consumer's and/or monopolist's welfare. Despite these drawbacks, the argument has strong intuitive appeal and is often invoked in welfare discussions of monopoly.

Similar reasoning is employed by Thompson (1968) in his attempt to refute the traditional underproduction hypothesis. To establish his result he simply demonstrated that the sum of individual's marginal valuations was less than the marginal cost of production. Thus, by producing one less unit of public good it is possible to increase the welfare of some subset of society without reducing the welfare of the remaining individuals. It is in this spirit that overproduction has been used.

Carrying this logic forward to the case at hand, it is easy to show that the sum of individuals' marginal valuations exceeds marginal cost of production; hence, underproduction can be said to exist. However, it is not necessarily legitimate to employ the marginal rate of substitution defined by (10) for this purpose. Consider those individuals whose consumption of B is less than the total amount produced. Their MRSs reflect the fact that they are denied access to existing units. Clearly such marginal valuations are inappropriate for purposes of deciding whether an additional unit of output should be produced. The relevant magnitude is the marginal valuation that would exist if they consumed total industry output.[12] These marginal valuations, in turn, depend upon the prices that we assume they pay for the extension of their consumption. In order to avoid the complications associated with income-effect feedbacks, we assume that the additional consumption is provided free of charge. As we shall demonstrate below, this particular assumption does not affect the qualitative nature of the conclusions to be derived. Let K individuals $(0 \leq K \leq N)$ comprise the class of consumers who, in equilibrium, are denied access to some part of the output of public goods. Let V^K be the sum of marginal valuations of this group, once individuals are allowed to extend, free of charge, their consumption to total output. For each of the $(N - K)$ individuals who do consume total output, we have by (10)

$$\frac{c}{N-K} \leq MRS^i. \qquad i = (N-K, N-K+1, \ldots, 1) \qquad (12)$$

---

[12] This assumes that the marginal utility of the public good is nonnegative. Otherwise, extension of consumption should continue only as long as marginal utility is positive.

Summing (12) over the $(N - K)$ individuals, we have

$$c \leq \sum_{i=1}^{N-K} MRS^i. \tag{13}$$

Now if $V^K > 0$, underproduction is seen to exist. That is, if the excluded group is willing to pay any positive sum for an additional unit of output, once they are allowed to consume total output, the traditional proposition holds. This appears to be a very weak condition, especially if we assume the marginal utility of the collective good to be nonnegative. In general, the presumption is that $V^K > 0$, irrespective of how the extension of public good consumption to the K individuals is financed. Of course, no such conclusion can be drawn if $K = 0$, the case where all individuals are identical. In this limiting case, (13) will hold as an equality. Our conclusion, therefore, is that underproduction is the general outcome of the competitive production of public goods. In no event will over production be the case.

V. Conclusion

This paper has explored the consequences of private provision of public goods when exclusion costs are zero, unit production costs are nondecreasing, entry into production and sale is free, and there is a lack of knowledge of specific individual preferences by firms. It was shown that a determinate nonzero level of production would result. However, unlike provision of private goods, the market outcome for public goods is unsatisfactory in two major respects: (1) some individuals are excluded from consuming all units of the public good; (2) the level of public good output is suboptimal. Hence, the orthodox position that private markets will underprovide for public goods for which exclusion is prohibitively costly can be extended to those for which exclusion costs are zero.

Before concluding, we must face the criticism that real-world examples of public goods produced in competitive markets are difficult to find. Thus, why study such cases? Several considerations can be brought to bear on this issue. (1) While it is true that under present arrangements examples are difficult to come by, we need not take such arrangements as given. Thus, one could imagine market provision of broadcasting and outdoor recreation facilities such as parks, police, and fire protection. Before rejecting private provision of such excludable services as an alternative, we must know some of its characteristics. (2) There is increasing pressure on state and local governments to consider user finance for many of their activities. The present analysis sheds light on the outcome of a decentralized self-supporting finance arrangement. (3) It is of inherent theoretical interest. It is interesting to know whether the problems posed by public goods stem from characteristics other than their inability to be produced in competitive markets. The present paper demonstrates that the lack of competitive markets for public goods is not solely responsible for deviations from optimality.

References

Bohm, Peter. "Estimating Demand for Public Goods: An Experiment." *European Econ. Rev.* 3 (1972): 111–30.

Buchanan, J. *The Demand and Supply of Public Goods.* Chicago: Rand-McNally, 1968.

Buchanan, J., and Kafoglis, M. Z. "A Note on Public Goods Supply." *A.E.R.* 53 (June 1963): 403–14.

Demsetz, Harold. "The Private Production of Public Goods." *J. Law and Econ.* 13 (October 1970): 292–306.

Ekelund, R., and Hulett, J. "Joint Supply, the Taussig-Pigou Controversy, and the Competitive Supply of Public Goods." *J. Law and Econ.* 16 (October 1973): 369–87.

Ganguly, Subrata K. "The Perfectly Competitive Production of Collective Goods: Comment." *Rev. Econ. and Statis.* 51 (November 1969): 478–79.

Johansen, Leif. *Public Economics.* Amsterdam: North-Holland, 1965.

Mishan, E. J. "The Postwar Literature on Externalities: An Interpretive Essay." *J. Econ. Literature* 9 (March 1971): 1–28.

Musgrave, Richard A. *The Theory of Public Finance.* New York: McGraw-Hill, 1959.

Olson, Mancur. *The Logic of Collective Action.* Cambridge, Mass.: Harvard Univ. Press, 1965.

Owen, Bruce M. "The Perfectly Competitive Production of Collective Goods Comment." *Rev. Econ. and Statis.* 51 (November 1969): 475–76.

Rodgers, James D. "The Perfectly Competitive Production of Collective Goods Comment." *Rev. Econ. and Statis.* 51 (November 1969): 476–77.

Samuelson, Paul A. "The Pure Theory of Public Expenditures." *Rev. Econ. a Statis.* 36 (November 1954): 387–89.

———. "Aspects of Public Expenditure Theories." *Rev. Econ. and Statis.* 40 (November 1958): 332–47.

———. "Pure Theory of Public Expenditure and Taxation." In *Public Economics,* edited by J. Margolis and H. Guitton. New York: American Elsevier, 1969.

Thompson, Earl A. "The Perfectly Competitive Production of Public Goods." *Rev. Econ. and Statis.* 50 (February 1968): 1–12.

———. "A Reply." *Rev. Econ. and Statis.* 51 (November 1969): 479–82.

Taxes, Spending, and Property Values: Supply Adjustment in a Tiebout-Oates Model

---

Matthew Edel

Queens College

Elliott Sclar

Brandeis University

This paper extends the capitalization approach of Wallace E. Oates to consider supply adjustment in a local public goods market of the sort hypothesized by Charles M. Tiebout. It is shown that Oates's single-period cross-section analysis-demonstrated demand conditions approximated Tiebout's hypotheses in a situation in which supply was not in long-run equilibrium. Analysis of capitalization of taxes, school expenditures, and road maintenance expenditures over five successive census periods in the Boston area indicates a move toward equilibrium in the market for schooling, but not in all other markets for local public goods.

I

Nearly 2 decades ago, Charles M. Tiebout (1956) proposed his well-known hypothesis that provision of local government services "reflects the preferences of the population more adequately than they can be reflected at the national level." In response to the arguments of Samuelson (1954) and Musgrave (1959) that no market test existed for the optimal level of (national) public goods output, Tiebout proposed a model under which a market analogue could lead to optimal expenditure on local public goods. The Tiebout model has since become a source of support for local

---

A preliminary version of this paper was presented at the annual meetings of the Western Economic Association, August 16, 1973. Research was funded by the Economics Division of the National Science Foundation and by the Center for the Study of Metropolitan Problems (Metro Center), National Institute of Mental Health. The comments of Ronald Grieson and the assistance of Eugene Kroch and Herman Thomas are acknowledged with thanks.

[*Journal of Political Economy*, 1974, vol. 82, no. 5]

© 1974 by The University of Chicago. All rights reserved.

fiscal autonomy, and is frequently cited in discussions of revenue sharing and in opposition to metropolitan consolidations. This has been particularly the case since Wallace E. Oates's (1969) publication of an "empirical study of tax capitalization and the Tiebout hypothesis," which is often interpreted as confirming the Tiebout model.

This paper extends Oates's capitalization approach to consider supply adjustment in the local public goods market. We argue that Oates restated Tiebout's model into a simpler question of whether households have preferences for a tax-service mix. This formulation does not consider whether the supply of public service combinations approaches the elasticity required for the model to yield Tiebout's analogue to the perfectly competitive solution. Oates verified that consumer demand conforms to Tiebout's hypothesis. However, this result was found only because supply was not in long-run equilibrium. If supply conditions in all markets had attained the competitive equilibrium, no capitalization would have been measured.

To investigate further the dynamics of an Oates-type market, we relate house values to local taxes and service delivery over different time periods for the Boston area. For schooling, there was a movement toward competitive equilibrium between 1950 and 1970. However, both continued tax capitalization despite the declining capitalization of school-expenditure differentials, and the continuing failure of highway-expenditure differentials to be capitalized indicate that supply conditions are not moving toward equilibrium in the markets for all public services. These empirical findings indicate that the Tiebout hypothesis cannot be considered as verified for local public service markets in general but only for some particular local service markets.

II

We begin by considering the manner in which local service benefits might be capitalized into land values, or location rents for sites in different tax-service districts, in a system characterized by the Tiebout model. In particular, we may assume that Tiebout's first five assumptions either are met directly in the system we are considering, or affect land values in a purely additive manner if they are unmet. That is, we assume (1) mobility and (2) knowledge among a population of utility-maximizing residents, (3) sufficient numbers of communities to insure competition among them and the availability of communities for each consumer's taste, (4) absence of restrictions on location due to employment needs (or, at least, differentiation among communities on grounds of accessibility to jobs that will enter into land values as a simple function of distance), and (5) absence of externalities between communities from public services.

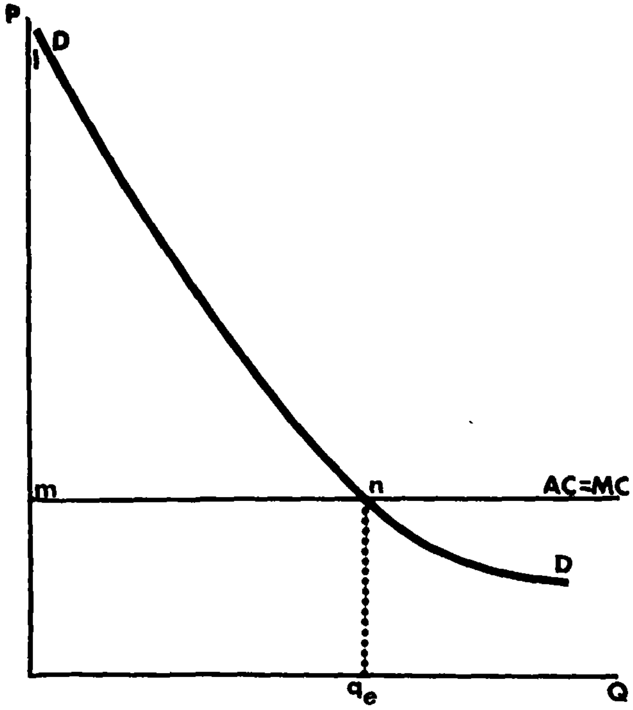


FIG. 1.—Supply and demand in a market for secular junior high school educations

Since we are assuming an analogue to a perfect market, point 3 implies that the residents of each town will be homogeneous with respect to their taste for each public service. Thus, for example, if education is provided, families without children, those with few children, and those with many children will be able to find towns fitting their needs and identical in all respects except for the provision of schooling. In addition, we begin with the assumption—which Tiebout approximates in the sixth and seventh conditions[1]—that constant returns to scale prevail in the provision of the service, or that, at least, equilibrium is possible with each town at a minimum-cost scale. For the empirical results presented later to hold it is also necessary to assume that funding is based on a local property tax rather than state or federal subsidies, which would add to benefits in some towns in a nonrandom manner.

In this situation, how will a market be represented in town provision of services? In figure 1, the cost of providing some public service (e.g., secular junior high school educations) is represented as $\$m$/unit (where a unit is, say, a student educated); *DD* is the demand curve for these junior high school educations. Since the average cost is passed on to consumers as a local tax in towns with these schools, it appears to them as a price. As many people move to towns with schools as are willing to pay a tax of *m* per average number of children per family. Since average cost, tax per pupil, and marginal cost are all the same, the market result is efficient in that marginal demand (marginal utility) is equated to marginal cost. As long as either there are true constant returns to scale, or the quantity demanded q_e is multiple of the optimal size, minimum cost is also insured.

---

[1] The actual assumptions used by Tiebout are (6) that for each pattern of consumer wants there is an optimal community size (which yields least costs), and (7) that communities try to attain the optimal size to minimize average costs.

In this case, the consumers of secular public junior high schools will reap a total consumer surplus of *lmn*. The schools will cover their costs. But will there be a corresponding producers' surplus to be capitalized into land values? Assuming a sufficient number of towns that anyone can find an otherwise acceptable location either with or without a secular junior high school, no resident would offer an after-tax rent above that attributable to nonschool-oriented attributes of the town. Apart from the additional tax paid for the schools, no resident would pay more than anyone locating in a school-less town to avoid the tax because of a lack of children or a preference for religious junior high schools (which may collect their average costs under the name of tuition fees rather than taxes). As a result, there is neither an increment left to capitalize in land values nor a reduction in ground values caused by the tax.

By contrast, we may see what occurs if the number of towns with schools is strictly limited. In figure 2, the assumption is made that only two towns have junior high schools, when it would require three towns to hold all of the people who would wish to use them, assuming prices were set equal to the average educational cost within those two towns. In this case, the effective supply curve is *mnsS'*, and the market-clearing price is *r*. Assuming the towns simply try to cover school costs with their taxes rather than making a profit for the public fisc, average taxes will be *m*, and a consumer's surplus of *mrlsn* would be available for those lucky enough to reside in those towns. However, the oversupply of would-be residents at price *m* would be expected to drive the price up to *r*. In this case, the area *mnsr* would represent the portion of consumer benefits captured by land rents. There would be, here, net capitalization of a portion of school benefits above real average costs. On the other hand, if a third town were then to open a school—at the same cost per pupil *m*—land value capitalization would disappear. The eventual situation would be as in figure 1, with the prior capitalization a temporary phenomenon due to the quicker approach to equilibrium by the first towns to open schools. Thus, if there are long-run constant returns to scale in service provision, an equilibrium can be approached as in the Tiebout model, but as that equilibrium is reached differential land values will be removed.

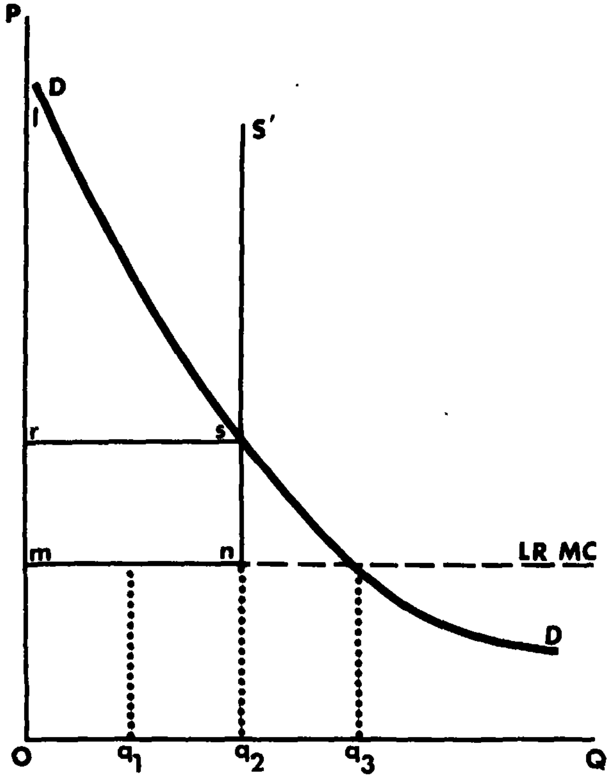


FIG. 2.—A market for schools with supply restricted to two towns

Relaxation of the assumption of constant returns alters the result only slightly. In figure 3, average and marginal costs rise with the number of users or the amount of the service provided. Such a situation may arise if production of the service uses scarce resources other than land (as in the case of health service provision if the supply of doctors and nurses is not perfectly elastic). If the service is sold at its marginal cost, as in a private goods market, the market will insure optimal resource allocation. There will be a producer's surplus *lrs*. This producer's surplus will take the form of land values, if one has to be a landowner to supply the service, as Grieson (1971) shows is the case for services of buildings. If, however, the service is paid for through taxation, hence sold on an average cost basis, the quantity purchased would be q_e, and while consumer's surplus would be greater producer's surplus would be eliminated. Nor would any of the consumer's surplus be capitalized into land values, for the prospective consumers would all be in equilibrium, paying only a price (the tax) equal to marginal utility (marginal demand).

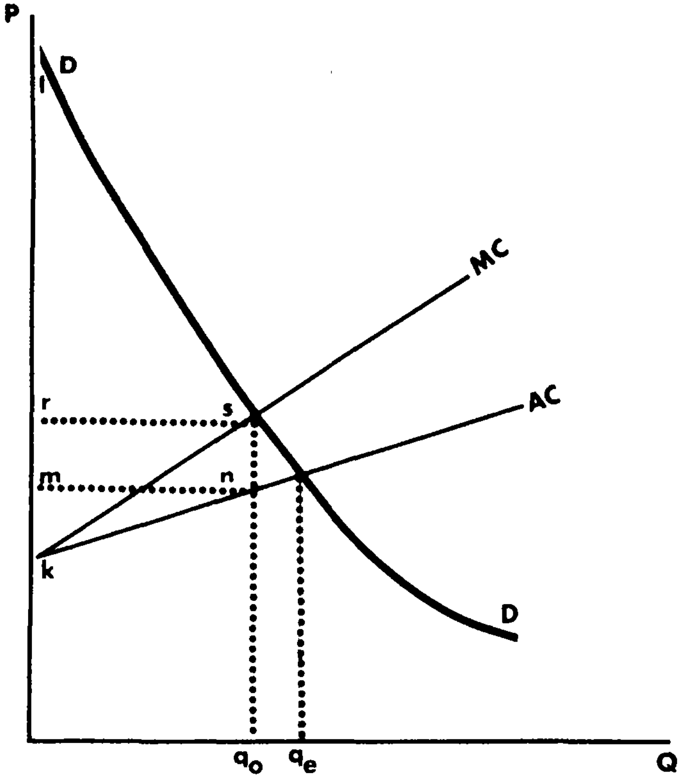


FIG. 3.—A market for medical facilities with doctors a scarce good

Capitalization will occur if some nonprice mechanism is used to limit demand to q_o, the optimal quantity at which demand intersects marginal cost. If, for example, zoning is used to limit entry into the school district, but taxes continue to cover average costs alone, then, for those consumers able to enter, marginal benefits will be at the level *r* while tax costs will equal *m*. The difference will be left as a pure consumer surplus (*mnrs* in addition to the original *lrs*) to families allowed in, if rationing (e.g., racial discrimination) is used to limit entry. However, if land is sold on a market, subject only to a density-zoning constraint, then the price per user must rise to *r*, and the area *mnrs* is left capitalized into land values. (This area is equal to the producer's surplus *krs* in the case of market sale of the service itself.)

In short, if only one service at a time is considered, the first five assumptions of the Tiebout model lead to a situation in which, at equilibrium, there is no capitalization effect. Taxes cover costs simply, and there is neither a net increment in land values from the presence of the service nor a decrease from the presence of the tax. Since the tax is a form of average-cost pricing of the service, the equilibrium may or may not be Pareto optimal. In the case of constant returns to scale, presence of capitalization indicates inefficiency. If there are decreasing returns to scale, the presence of positive capitalization is compatible with efficiency (although it does not guarantee that a situation of efficiency exists).[2] In the case of increasing returns to scale, by a similar argument, the provision of efficient service levels financed out of average-cost local taxation would involve negative capitalization.[3] But in these last two cases, efficiency (if it were to exist) would require methods of limiting or attracting entrants to the community through methods not foreseen in the Tiebout model. Zoning might, in this case, make a situation either more or less optimal. To show that there is capitalization in a model with only one public service thus fails to show that a Tiebout mechanism has assured either equilibrium or efficiency.

The situation is more complex if different public services are present within a system of competing towns (see Buchanan and Goetz 1972). In this case, if some services are in equilibrium and some not in equilibrium, different total levels of taxation may be associated with different levels of land values, and presence of services whose provision and financing are in equilibrium may appear to be associated with higher land values. Suppose one service (let us say schooling) is provided by a Tiebout system that is in equilibrium, but there are other services which are not provided in equilibrium. The difference between marginal demand and average cost for these other services will be capitalized. If, however, a comparison of towns, using regression analysis, treats the total tax rate as a single variable, then the net noncapitalization of school taxes and expenditures may appear as offsetting positive capitalization of the expenditures and negative capitalization of the taxes. That is, by comparing the school taxes with other taxes that are unmatched by benefits, the two elements of capitalization are conceptually separable. But if average-cost taxes and marginal benefits are equal in the provision of all services, there will be no way of making this separation by comparison of the towns. If the Tiebout model held perfectly (at least in its first five assumptions) for all services, a regression analysis relating land values to taxes and service levels, holding other background conditions constant, would show no capitalization of any taxes or services.

---

[2] The Pareto optimal situation in this case may involve a more unequal distribution of benefits than a case of "inefficient" oversupply of the service where nobody is excluded by nonmarket means. The problems involved in applying the Tiebout model to cases of income redistributing public services are explored by Rothenberg (1969, 1970) and by Miller and Tabb (1973).

[3] One parallel to this is Goldberg's (1970, 1972) demonstration that a transportation system under certain assumptions is most efficient if it minimizes land values (see also Edel 1971).

III

The preceding discussion has indicated that a world in which the conditions of the Tiebout model were met perfectly would see no tax and local expenditure capitalization into property values. Oates sought to demonstrate the "relevance to actual behavior" of the Tiebout model by a demonstration of capitalization using New Jersey data for 1960. His linking of an observable quantity, property value, to the Tiebout model is a fruitful step in public finance theory. However, the version of the Tiebout model which Oates tests is oversimplified.

Tiebout, it will be remembered, presented his theory as a response to the Musgrave-Samuelson agreement that "no 'market type' solution exists to determine the level of expenditures on public goods. Seemingly we are faced with the problem of having a rather large portion of our national income allocated in a 'non-optimal' way when compared with the private sector" (Tiebout 1956). Indeed, his very title was a play on Samuelson's 1954 article. Tiebout's paper sought to show that, at least conceptually (under what he admitted were "extreme assumptions"), a "market type" solution might allow the optimal provision of public services out of local taxes.

Oates, on the other hand, simplifies the Tiebout model into a statement about consumer demand only, with no supply side made explicit, and no optimality issue raised: "Very simply, Tiebout's world is one in which the consumer 'shops' among different communities offering varying packages of local public services and selects as a residence the community which offers the tax-expenditure program best suited to his tastes" (Oates 1969, p. 957). His results, therefore, are never compared with the results predicted from a full Tiebout model. They are, rather, a demonstration that taxes and services have some effect on consumer preferences for town of residence, which is reflected in house prices. The tax and service capitalization which Oates finds would not occur if consumers were completely indifferent to services and taxes, had no information about their differences among towns, or succeeded in passing all local taxes on completely to the federal government in reduced income taxes. Oates eliminates these possibilities, demonstrating the individual economic rationality of the home buyer.[4] But his results also would not occur if the pure theory as presented by Tiebout were a complete description for all local service markets.

In the specific case of school expenditures, Oates calculates, using "typical" values, that the amount of decrease in home values from capitalization of approximately two-thirds of a 1 percent tax rate increase would roughly equal the increase in home prices from the spending of the money raised on an average school. The calculation is presented "as indicating no more than orders of magnitude" and as a suggestion that, tax and expenditure capitalization being equal, the market for schooling may be in Tiebout's equilibrium. However, if the interest rate used for capitalization or the two-thirds rate of capitalization were different, the lack of net capitalization of schooling would not hold. A Tiebout equilibrium for education is thus demonstrated only on the basis of strong assumptions, while that at least one sector of town expenditures must be out of equilibrium is conclusively demonstrated by the finding of any capitalization. A similar result follows from the similar finding of capitalization by Heinberg and Oates (1970) using 1960 Massachusetts data.

---

[4] Other studies have used analyses of price responsiveness to demonstrate that peasants are "rational" (for discussion see Edel 1970).

Oates concludes from his analysis: "These results appear consistent with a model of the Tiebout variety in which rational consumers weigh (to some extent at least) the benefits from local public services against the cost of their tax liability in choosing a community of residence; people do appear willing to pay more to live in a community which provides a high-quality of public services (or in a community which provides the same program of public services with lower tax rates)" (Oates 1969, p. 968). This conclusion is warranted, but it is incomplete. "A model of the Tiebout variety" as used by Oates does not consider the elasticity or inelasticity of supply and hence whether the welfare-economic implications of Tiebout's original model hold. By introducing the analysis of capitalization into the discussion of Tiebout's model, Oates has taken a significant step. However, it is necessary to go beyond his single-period measure of capitalization to examine changes in capitalization during supply adjustment in order to determine whether a Tiebout equilibrium is being approached. In the following section, Oates's capitalization approach is used to examine such a historical process.

IV

Examination of data on house values and public finance for a group of towns in the Boston metropolitan area over several decades allows such an examination of the supply-adjustment capitalization process.[5] Boston area data provide an appropriate test because of the extreme dependence of towns in that vicinity on the local property tax for financing services. In Massachusetts, 54 percent of local expenditures were financed from this tax as of 1970, as compared to a 40 percent national average. Highway and educational maintenance expenditures, the two funding categories to be considered in detail, have particularly low levels of outside aid. The

---

[5] This study is part of a longer historical examination of factors influencing Boston-area property values.

III

The preceding discussion has indicated that a world in which the conditions of the Tiebout model were met perfectly would see no tax and local expenditure capitalization into property values. Oates sought to demonstrate the "relevance to actual behavior" of the Tiebout model by a demonstration of capitalization using New Jersey data for 1960. His linking of an observable quantity, property value, to the Tiebout model is a fruitful step in public finance theory. However, the version of the Tiebout model which Oates tests is oversimplified.

Tiebout, it will be remembered, presented his theory as a response to the Musgrave-Samuelson agreement that "no 'market type' solution exists to determine the level of expenditures on public goods. Seemingly we are faced with the problem of having a rather large portion of our national income allocated in a 'non-optimal' way when compared with the private sector" (Tiebout 1956). Indeed, his very title was a play on Samuelson's 1954 article. Tiebout's paper sought to show that, at least conceptually (under what he admitted were "extreme assumptions"), a "market type" solution might allow the optimal provision of public services out of local taxes.

Oates, on the other hand, simplifies the Tiebout model into a statement about consumer demand only, with no supply side made explicit, and no optimality issue raised: "Very simply, Tiebout's world is one in which the consumer 'shops' among different communities offering varying packages of local public services and selects as a residence the community which offers the tax-expenditure program best suited to his tastes" (Oates 1969, p. 957). His results, therefore, are never compared with the results predicted from a full Tiebout model. They are, rather, a demonstration that taxes and services have some effect on consumer preferences for town of residence, which is reflected in house prices. The tax and service capitalization which Oates finds would not occur if consumers were completely indifferent to services and taxes, had no information about their differences among towns, or succeeded in passing all local taxes on completely to the federal government in reduced income taxes. Oates eliminates these possibilities, demonstrating the individual economic rationality of the home buyer.[4] But his results also would not occur if the pure theory as presented by Tiebout were a complete description for all local service markets.

In the specific case of school expenditures, Oates calculates, using "typical" values, that the amount of decrease in home values from capitalization of approximately two-thirds of a 1 percent tax rate increase would roughly equal the increase in home prices from the spending of the money raised on an average school. The calculation is presented "as indicating no more than orders of magnitude" and as a suggestion that, tax and expenditure capitalization being equal, the market for schooling may be in Tiebout's equilibrium. However, if the interest rate used for capitalization or the two-thirds rate of capitalization were different, the lack of net capitalization of schooling would not hold. A Tiebout equilibrium for education is thus demonstrated only on the basis of strong assumptions, while that at least one sector of town expenditures must be out of equilibrium is conclusively demonstrated by the finding of any capitalization. A similar result follows from the similar finding of capitalization by Heinberg and Oates (1970) using 1960 Massachusetts data.

---

[4] Other studies have used analyses of price responsiveness to demonstrate that peasants are "rational" (for discussion see Edel 1970).

Oates concludes from his analysis: "These results appear consistent with a model of the Tiebout variety in which rational consumers weigh (to some extent at least) the benefits from local public services against the cost of their tax liability in choosing a community of residence; people do appear willing to pay more to live in a community which provides a high-quality of public services (or in a community which provides the same program of public services with lower tax rates)" (Oates 1969, p. 968). This conclusion is warranted, but it is incomplete. "A model of the Tiebout variety" as used by Oates does not consider the elasticity or inelasticity of supply and hence whether the welfare-economic implications of Tiebout's original model hold. By introducing the analysis of capitalization into the discussion of Tiebout's model, Oates has taken a significant step. However, it is necessary to go beyond his single-period measure of capitalization to examine changes in capitalization during supply adjustment in order to determine whether a Tiebout equilibrium is being approached. In the following section, Oates's capitalization approach is used to examine such a historical process.

IV

Examination of data on house values and public finance for a group of towns in the Boston metropolitan area over several decades allows such an examination of the supply-adjustment capitalization process.[5] Boston area data provide an appropriate test because of the extreme dependence of towns in that vicinity on the local property tax for financing services. In Massachusetts, 54 percent of local expenditures were financed from this tax as of 1970, as compared to a 40 percent national average. Highway and educational maintenance expenditures, the two funding categories to be considered in detail, have particularly low levels of outside aid. The figure for highway expenditures, it should be noted, does not include construction of new roads or highways, for which outside funding is considerable, but only their upkeep, which is a considerable expense in a snow-affected area like New England.

---

[5] This study is part of a longer historical examination of factors influencing Boston-area property values.

Using a somewhat different sample from the Massachusetts data than Heinberg and Oates, we estimated regressions of a somewhat similar form to theirs, for the 5 census years beginning with 1930. These regressions use six distance variables (latitudinal distance from downtown Boston, longitudinal distance, and their squares and cubes) rather than a single accessibility variable. A variable for per-square-mile highway maintenance expenditures is a second form of public expenditure, along with school expenditures per pupil, whose capitalization into house values is tested. Density and percentage of owner occupiers among residents of a town are used to capture amenity factors not explained by distance. The home-ownership term may also capture the shifting of the property tax through federal income tax deductions. Tax rates are nominal rates for 1930–60, and both nominal and equalized rates are employed in separate regressions for 1970.

Estimated coefficients appear in table 1 and are superficially similar to Oates's findings for the postwar period. Tax-rate capitalization is negative in all years and significant for 1940, 1950, and 1970. School expenditures per pupil are capitalized positively in all 3 postwar census years. However, for 1960 and 1970, the effect is not statistically significant at the .01 level and is quantitatively smaller than for 1950. For 1930 and 1940 the school-expenditure term is not significant, and indeed the sign is negative. The highway-maintenance expenditures are not significantly capitalized in any period.

The regressions appear to show somewhat weaker capitalization than Oates's examples. Assuming a $21,000 house with a tax rate of $50 per $1,000 (a typical figure 5 miles north or south of the city in the 1970 regressions), a $10 reduction in the effective tax rate would amount to a $210 reduction in taxes. Capitalized at an 8 percent interest rate, this would amount to a $2,625 increase in value, about twice the observed increase given by the regression estimate. In other words, tax capitalization in this case would be about 50 percent, rather than the two-thirds found by Oates. Perhaps more important than this difference is the declining payoff in house values from each dollar per pupil spent on schools: the capitalization falls from almost $24 in 1950 to under $9 in 1960 and about $3 in 1970. At first glance, this might appear to show that Bostonians are less concerned with education and taxes than are Oates's New Jersey residents and that they are becoming less concerned with schools over time.

An alternative explanation suggests an approach, over time, to a Tiebout equilibrium in the provision of schooling. According to Hirsch

TABLE 1

REGRESSION ESTIMATES OF EFFECT OF TAXES, SCHOOL AND ROAD EXPENDITURES, AND OTHER VARIABLES ON HOUSE PRICES

Year	Tax Rate ($/$1,000)	Density (Pop./Sq. Mile)	Owner Occupied	School Expenditures ($/Pupil)	Highway Maintenance ($/Sq. Mile)	R^2 (n)
1930	−503	0.050	n.a.	−2.31	−0.006	.349
	(1.31)	(0.73)		(0.28)	(0.44)	(78)
1940	−651	−0.015	−5.01	−3.55	−5.01	.355
	(2.49)	(0.33)	(0.40)	(0.67)	(0.40)	(77)
1950	−811	−0.017	52.48	23.68	−0.009	.648
	(3.03)	(0.17)	(1.99)	(4.08)	(0.49)	(77)
1960	−126	−0.400	0.001	8.77	0.008	.376
	(0.54)	(2.56)	(0.03)	(1.59)	(0.54)	(73)
1970	−43.3	−0.374	86.67	2.99	0.012	.546
	(2.18)	(1.49)	(1.41)	(1.48)	(0.74)	(65)
1970* ...	−139.0*	−0.211	127.87	3.25	0.023	.551
	(2.06)	(0.59)	(2.10)	(1.60)	(1.39)	(64)

NOTE.—These are values of *b* from regressions in which latitudinal and longitudinal distance and their squares and cubes were also independent variables. House value is measured in dollars from census estimates of town medians. (Values of *t* are in parentheses.)

* Using equalized tax rates (other regressions use nominal tax rates).

(1968), schooling is a service which can be provided efficiently by relatively small units. Economies of scale exist only up to sizes of perhaps a few hundred pupils for elementary schools and 1,700 for secondary schools. These sizes are within the population range of most Boston area communities; the few that are too small to support high schools can easily form consolidated high school districts. There are also sufficient independent towns in each geographic region of the Standard Metropolitan Statistical Area to allow consumer choice between towns with expensive schools and towns with low school expenditures which attract childless individuals or parents preferring private and parochial schools. (Among the higher-income communities, the comparison between Brookline and Belmont is most obvious.) The conditions cited by Tiebout as assumptions in his model might thus seem to hold for schools. This explanation is consistent with the 1970 regressions: assuming each family, on the average, has three children of school age, $1.00 per pupil spent by the schools would represent a $3.00 benefit to the family (if taxes and other community and housing characteristics are held constant). The $3.00 increase in house values is thus an approximation to a simple capitalization of the benefits, which would be matched exactly by an offsetting loss of $3.00 *if* families paid average cost through taxes and if taxes were fully capitalized. In practice, the equilibrium is not exactly obtained: tax capitalization appears only partial, the average (although perhaps not the marginally moving) family has less than three children, and some school expenditures are borne by state and federal aid rather than local property taxes. Nonetheless, comparing the 3 years—1950, 1960, and 1970—it is the low capitalization figure for school expenditures in 1970 that is approximately consistent with a Tiebout equilibrium.

But what, then, of the higher capitalization in benefits for 1950 and 1960? It would appear that these represent higher returns due to a shortage of schools. A situation obtains such as that in figure 2 rather than figure 1; due to limited supply the marginal demand for schooling is far above the average cost. Beginning in the early postwar period, that is to say, there appears to have been a shortage of adequate quantities and qualities of schools, given the baby boom, the rising need for technical skills in the labor force, and the much-noted shortage of teachers and school facilities. In this situation, parents would pay far more than the actual average cost of schooling through increased house prices in order to enter a community with what they considered to be adequate educational facilities. By 1960, the shortage was somewhat reduced from the 1950 level, and capitalization of per-pupil expenditures had declined somewhat. By 1970, there appears to have been very little markup, if any, left. The comparison of the three periods thus seems to be much like the movement from a short-run to a long-run equilibrium in Marshallian economics.[6]

Tiebout's model thus seems to have been a fair representation of the schooling market in the Boston area by 1970. Oates, measuring returns in 1960 both for New Jersey and (with Heinberg) for Massachusetts, would seem to have found greater returns because he took a cross-section view before equilibrium was achieved. His speculation that "it could be . . . the negative association we have observed between property taxes and home values is primarily a short-run phenomenon, which would disappear over a longer period of time" (Oates 1969, p. 967) has thus been borne out. Indeed the logic of the Tiebout model indicates this should occur in an approach to equilibrium

The figures in table 1 also indicate, however, that while the provision of schooling may have approached a Tiebout equilibrium, the same could not have been true for all other local services. In 1970 tax rates (whether viewed in real or nominal terms) significantly affected house values, even though they were not 100 percent capitalized. What is more, local road-maintenance expenditures were not positively capitalized to a statistically significant extent in any period, although some positive effect is found for 1970. If all expenditures were in a Tiebout equilibrium, taxes should not have been capitalized, while, since taxes were capitalized, road maintenance should have had positive effects on house values if it had positive benefits from a consumer viewpoint. Thus not all markets can be in Tiebout equilibria, and the road-maintenance market appears likely as a candidate to be out of equilibrium.

---

[6] The explanation for the negative capitalization of school expenditures for 1930 and 1940 may be that, because of the depression, some communities were left with school average costs above marginal benefits, which they were unable to reduce.

Road expenditures are as clear an example of violation of Tiebout's conditions as schools are a conforming example. Neither are there clear constant returns to scale, nor is there an absence of spillovers. Indeed, in the Boston area, highway-maintenance expenditures appear mainly related to the flow-through of traffic and population density. Highway expenses in dollars per square mile (HXP) for 1964 are explained by the regression:

$$\mathrm{HXP} = 34857 + \underset{(10.01)}{31.62\ \mathrm{FLOW}} + \underset{(10.94)}{11.27\ \mathrm{RES}}\ (R^2 = .784),$$

where FLOW = commuter flow-through per square-mile area, and RES = residual population density estimated from:

$$\mathrm{DENSITY} = 2567 + \underset{(10.29)}{4.00\ \mathrm{FLOW}}\ (R^2 = .631).$$

Addition of a term for median per-family income has no significant effect. Inner-ring communities must maintain roads through which out-of-town commuters pass although their residents get few benefits. Only a few road expenses, such as frequency of snow removal in winter on side streets, are really optional expenses that may rise with income or local preference. Thus, differences in road-maintenance expenditures do not represent a response to demand differences, as in the Tiebout model, and are not capitalized.[7] The same is true for some other road-related services such as traffic-police expenditure and contributions to the Metropolitan District Commission highway system.

Thus, despite the movement toward a Tiebout-type equilibrium for schooling, there appear to be public service markets (including road maintenance, for one) which do not show movement toward such an equilibrium. The Tiebout hypothesis appears to hold for some public service markets. But, since it does not hold for all of them, prudence (reinforced by the general theory of the second best) would caution against acceptance of the notion that a "market-type" mechanism of the Tiebout sort has beneficial efficiency properties if applied to all public services taken together, as compared with any possible alternative systems of public-expenditure determination for metropolitan areas.

References

Buchanan, J. M., and Goetz, C. J. "Efficiency Limits of Fiscal Mobility: An Assessment of the Tiebout Model." *J. Public Econ.* 1 (April 1972): 25–43.

Edel, M. "Innovative Supply: A Weak Point in Economic Development Theory." *Soc. Sci. Information* 9 (June 1970): 9-40.

---

[7] Further discussion of commuter-related expenses is found in Sclar (1971).

Edel, M. "Land Values and the Costs of Urban Congestion: Measurement and Distribution." *Soc. Sci. Information* 10 (December 1971): 7-36.

Goldberg, M. A. "Transportation, Urban Land Values, and Rents: A Synthesis." *Land Econ.* 46 (May 1970): 153-62.

———. "An Evaluation of the Interaction between Urban Transport and Land Use Systems." *Land Econ.* 48 (November 1972): 338-46.

Grieson, R. "The Economics of Property Taxes and Land Values." Massachusetts Inst. Tech., Department of Economics Working Paper, 1971.

Heinberg, J. D., and Oates, W. E. "The Incidence of Differential Property Taxes on Urban Housing." *Nat. Tax J.* 23 (March 1970): 92-98.

Hirsch, W. Z. "The Supply of Urban Public Services." In *Issues in Urban Economics*, edited by H. S. Perloff and L. Wingo, Jr. Baltimore: Johns Hopkins Univ. Press, 1968.

Miller, S., and Tabb, W. K. "A New Look at a Pure Theory of Local Expenditures." *Nat. Tax J.* 26 (June 1973): 161-76.

Musgrave, R. A. *The Theory of Public Finance*. New York: McGraw-Hill, 1959.

Oates, W. E. "The Effects of Property Taxes and Local Spending on Property Values: An Empirical Study of Tax Capitalization and the Tiebout Hypothesis." *J.P.E.* 77 (November/December 1969): 957-71.

Rothenberg, J. "Strategic Interaction and Resource Allocation in Metropolitan Intergovernmental Relations." *A.E.R.* 59 (May 1969): 495-503.

———. "Local Decentralization and the Theory of Optimal Government." In *The Analysis of Public Output*, edited by J. Margolis. New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1970.

Samuelson, P. A. "The Pure Theory of Public Expenditures." *Rev. Econ. and Statis.* 36 (November 1954): 387-89.

Sclar, E. "The Public Financing of Transportation in the Greater Boston Area." Ph.D. dissertation, Tufts Univ., 1971.

Tiebout, C. M. "A Pure Theory of Local Expenditures." *J.P.E.* 65 (October 1956): 416-24.

Short-Run and Long-Run Equilibrium for a Small Open Economy

Wolfgang Mayer

University of Cincinnati

This paper deals with a small open economy in which capital goods employed by a given industry are temporarily immobile and can be reallocated only gradually over an extended period of time. The short-run responses of employment, factor returns, and output levels to exogenous changes in commodity prices and endowments, as well as the following time-absorbing adjustment process of the endogenous variables to long-run equilibrium, are analyzed. In addition, it is argued that this short-run theory provides a better explanation of factor-owner reactions to trade policies than conventional long-run international trade theory.

I. Introduction

In many areas of economic theory one distinguishes between the short run, a time period so short that usage of certain inputs cannot be changed, and the long run, a time period long enough to allow for changes in employment of all inputs. The standard models in the theory of international trade, however, do not operate with similar distinction.[1] It

I am much indebted to Jacob Frenkel for his detailed comments, in particular for suggesting the paper's title and for referring me to the works of Haberler (1936) and Jones (1971). I should also like to thank Horst Herberg for his comments. Any errors and shortcomings of the paper are, of course, my sole responsibility. Some conclusions reached in this paper have independently been arrived at by Jorge Marquez-Ruarte in "An Elementary Framework of Inter-Run Analysis," mimeographed (University of Chicago, October 1973), a paper which the associate editor of this *Journal*, Robert Barro, made available to me. No attempt has been made to change my original version on account of this paper.

[1] This is somewhat surprising, since Haberler (1936), in his discussion of international trade theory, already distinguishes between the short and the long run. In fact, he points out that in the short run, when most factors are specific to a given industry, the production-possibility frontier's curvature will be more concave to the origin than in the long run (pp. 178–79). We should add that a long-run concept of a quite different nature has been used by Atsumi (1971). He develops a dynamic trade model in which allowance is made for capital accumulation and population growth.

[*Journal of Political Economy*, 1974, vol. 82, no. 5]
© 1974 by The University of Chicago. All rights reserved.

simply is assumed that factors of production are freely mobile within a country and that they are allocated in such a way that the value of a given factor's marginal product is the same in all industries. Clearly, this assumption implies that these models are "long run" in nature.

Two main results of this "long-run" trade theory are the Rybczynski and Stolper-Samuelson theorems. The former is a powerful tool in explaining differences in production patterns between countries with different factor endowments. The latter is helpful in explaining factor-return differentials between countries whose final commodity prices differ because of tariffs, transportation costs, or other forms of trade impediments. In both cases, the explanations are based on comparisons of long-run equilibria. The two theorems are, however, used for other purposes as well. The Rybczynski theorem is applied in attempts to determine the impact of endowment changes—through immigration, capital transfers, or other means—on factor returns and output levels. The Stolper-Samuelson theorem is applied to predict changes in factor rewards as tariffs are imposed or transportation costs change. In both cases, the theorems reveal what eventually is going to happen after enough time has elapsed for reestablishing an optimal, long-run factor-allocation pattern. While there is much value to such long-run predictions, it also is quite obvious that a very substantive vacuum is left by the two theorems. They neither answer any questions concerning the immediate impact of changes in endowment or commodity prices on the economy, nor do they deal with the adjustment of variables between long-run equilibria.

The purpose of our paper is to fill this vacuum. We are going to set up a simple general-equilibrium model and study the short-run effects of varying commodity prices and endowments on factor returns and output levels. Such a short-run theory is based on the assumption that not all factors can move freely between industries as incentives for reallocation arise; some factors are temporarily locked in and cannot be adjusted instantaneously. Only over a longer period of time, if incentives for reallocation continue, the initially locked-in factors will become mobile, and gradual adjustment to long-run equilibrium can take place. Hence, our analysis will not only deal with short-run equilibria, but also with tracing out the adjustment path for factor returns and output levels between different long-run equilibria.

Our results should be of more than theoretical interest. They should help us understand some very important real-world phenomena associated with international trade policies. For example, they illustrate why the liberalization of immigration restrictions is strongly opposed by workers and much favored by capital owners. They also make clear why workers and capital owners of a given industry share a strong common interest when it comes to tariff protection of their industry. Since people work, consume, and earn income in the short run, it is not surprising that such attitudes of factor owners toward trade policies cannot be explained by any long-run theory, such as we find in the standard models of international trade theory.[2] Only a short-run theory can provide meaningful answers to short-run questions.

In the next section, we specify the conditions for short- and long-run equilibrium and compare slope and curvature of the long-run with those of the short-run transformation curve. In Section III, we first study the short-run responses of factor returns and commodity outputs to price changes and then discuss the adjustment path of the variables to long-run equilibrium. Section IV deals with the same questions as the preceding section, except that in Section IV a change in labor endowment is the disturbing force.

II. Short-Run General Equilibrium

Let us make the usual assumptions of a two-commodity, two-factor economy which is incompletely specialized, in which factors are fully employed, and where perfect competition is prevailing. Factor endowments and final commodity prices are exogenously given. Since the model is a simple two-by-two barter model, only relative prices matter; accordingly, for the remaining analysis the first commodity is chosen as the numeraire. Industry production functions are represented by

$$X_j = X_j(L_j, K_j), \qquad j = 1, 2 \tag{1}$$

where X_j denotes output of commodity j, and L_j, K_j stand for the amounts of labor and capital, respectively, employed by industry j. The functions are assumed to be homogeneous of degree 1, twice differentiable, and strictly quasi-concave; furthermore, $X_{L_j} > 0$ and $X_{K_j} > 0$.

Long-run (LR) equilibrium is established if each factor's marginal product, evaluated in terms of the first commodity, is the same in both industries,

$$\begin{aligned} X_{L_1} &= pX_{L_2}, \\ X_{K_1} &= pX_{K_2}, \end{aligned} \tag{2}$$

and both factors are fully employed,

$$\begin{aligned} L_1 + L_2 &= \bar{L}, \\ K_1 + K_2 &= \bar{K}. \end{aligned} \tag{3}$$

---

[2] According to the Rybczynski theorem, the inflow of additional labor has no impact on factor returns as long as the economy remains nonspecialized. The Stolper-Samuelson theorem tells us that a protective tariff will raise returns on this factor, which is used relatively intensively in the production of the protected commodity and lower returns on the other factor.

In the preceding equations, p denotes the relative price of the second in terms of the first commodity, and $\bar{L}$ and $\bar{K}$ measure the economy's endowment with labor and capital. With the information contained in equations (1)–(3), we now can trace out the LR transformation locus, denoted by *FF* in figure 1. The locus reflects the economy's production potential, given factor endowments, technology constraints, and no impediments to the optimal allocation of factors. As is well known, the assumptions of perfect competition and production functions homogeneous of degree 1 imply that the LR transformation curve's slope is

$$(dX_2/dX_1)^{\mathrm{LR}} = -(1/p), \tag{4}$$

and that its curvature is convex outward:[3]

$$(d^2X_2/dX_1^2)^{\mathrm{LR}} \leq 0. \tag{5}$$

In the short run (SR), it is assumed, only labor is perfectly mobile between industries.[4] Capital, once installed in a given industry, cannot be transferred instantaneously to uses in the other industry, even though capital returns may differ between industries.[5] Consequently, SR equilibrium is described by the condition that labor's marginal product is the same in both industries,

$$X_{L_1} = pX_{L_2}, \tag{6}$$

and that the following resource constraints are satisfied:

$$\begin{aligned} L_1 + L_2 &= \bar{L}, \\ K_1 &= \bar{K}_1, \\ K_2 &= \bar{K}_2, \end{aligned} \tag{7}$$

where $\bar{K}_j$ is the optimal amount of capital employed by industry j in accordance with a predetermined LR equilibrium. For example, we may choose a point on the *FF* locus, such as A, which represents a LR equilibrium position corresponding to the relative price $(1/p)^A$. The optimal factor allocations, based on conditions (2) and (3), are K_1^A, K_2^A, L_1^A, and L_2^A. Using A as a reference point, we now can derive the SR transformation curve, denoted by *aa*. Along *aa*, the capital allocation is the same as at point A, namely, K_1^A, K_2^A, but the labor allocation varies in order to satisfy criterion (6). In contrast to the *FF* locus, which describes the economy's ultimate production potential, the *aa* locus expresses the

---

[3] See, for example, Lancaster (1968).

[4] The assumption of short-run immobility of a certain factor implies that this factor temporarily is specific to a given industry. Hence, our short-run model is basically the same as Jones's (1971) three-factor, two-commodity model, in which each industry uses one factor that is specific to it.

[5] Individual firms within a given industry are still able to choose their optimal capital-labor combinations.

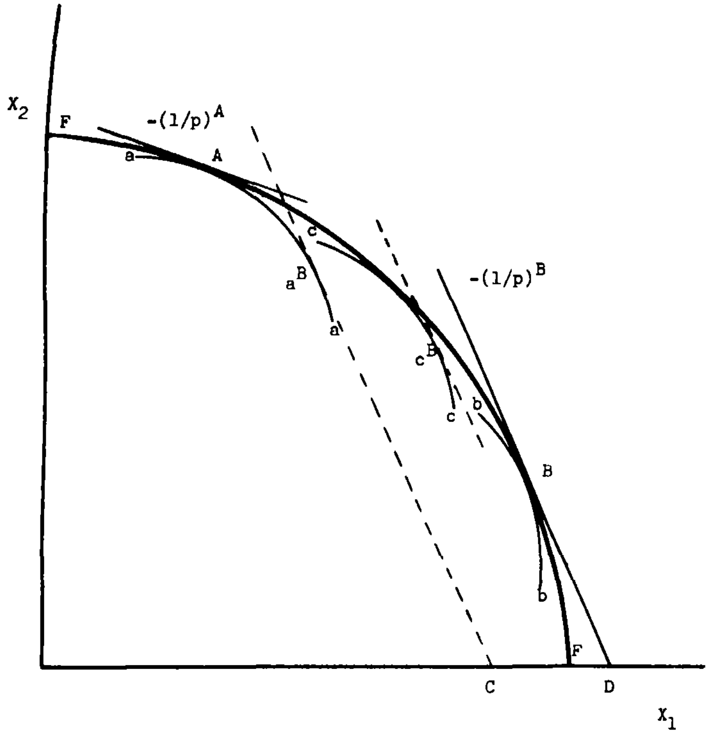


Fig. 1

economy's SR ability to adjust to price changes. If, for instance, the final commodity price ratio changes to $(1/p)^B$, new SR and LR equilibria indicated by points a^B and B, respectively, are established. Actual production will change to a^B, where the price line is tangent to the *aa* locus. Point B, on the other hand, only refers to future output combinations that will become attainable after the passage of sufficient time to allow for optimal reallocation of capital. Stated differently, the price change induces an instantaneous change of outputs from A to a^B, and then it brings about a time-dependent adjustment process starting at a^B and terminating at B. The SR equilibrium, a^B, is characterized by a temporary suboptimal allocation of capital, which is responsible for an income loss, measured in terms of commodity 1 by the distance *CD* in figure 1.

In the diagram, the SR equilibrium point a^B was determined by drawing the price line tangent to the *aa* locus. That a^B is, in fact, the SR equilibrium can be seen after we take (1), differentiate it totally for both industries while holding $\bar{K}_j$ constant, and set dL_2 equal to $-dL_1$. As we divide the expression for dX_2 by that for dX_1 and use (6) to satisfy the SR conditions we obtain:

$$(dX_2/dX_1)^{SR} = -(X_{L_2}/X_{L_1}) = -(1/p). \tag{8}$$

This proves that, at the point of equilibrium, the slope of the SR and LR transformation curves is the same. Their respective curvatures, however, are not identical. Since from a LR perspective all points on *aa*, except for *A*, represent suboptimal factor allocations, *aa* must lie inside *FF* except for point *A* itself, where *aa* is tangent to *FF*. Furthermore, the *aa* locus must be strictly convex outward.[6]

III. Short-Run Responses to Price Changes and the Adjustment to Long-Run Equilibrium

Exogenous price changes of final commodities affect factor returns and industry output levels. In this section, we are going to examine these responses in the SR, compare them with the LR results, and describe the adjustment path from the initial to the new LR equilibrium.

First, we will deal with the SR responses of labor and capital returns, both of which are evaluated in terms of commodity 1, and which for industry j are denoted by w_j and r_j, respectively. Given these definitions of w_j and r_j and the specification of the SR equilibrium conditions, as stated in (6) and (7), we obtain a system of four equations:

$$\begin{aligned} X_{L_1}(L_1, \bar{K}_1) &= pX_{L_2}(\bar{L} - L_1, \bar{K}_2), \\ X_{L_1}(L_1, \bar{K}_1) &= w, \\ X_{K_1}(L_1, \bar{K}_1) &= r_1, \\ pX_{K_2}(\bar{L} - L_1, \bar{K}_2) &= r_2. \end{aligned} \tag{9}$$

The endogenous variables are L_1, w, r_1, and r_2, which are to be evaluated at the initial SR equilibrium point. After differentiating (9) totally, holding labor endowment and capital allocations constant, we can solve for the percentage change in factor returns as a function of the percentage

---

[6] Total differentiation of (8) with respect to X_1 yields

$$(d^2X_2/dX_1^2)^{SR} = -(1/X_{L_1}^2)[X_{L_1}X_{L_2L_2}(dL_2/dL_1) - X_{L_2}X_{L_1L_1}](dL_1/dX_1).$$

Substituting $(dL_2/dL_1) = -1$ and $(dL_1/dX_1) = (1/X_{L_1})$, the above expression can be reduced to:

$$(d^2X_2/dX_1^2)^{SR} = (1/X_{L_1})^3(X_{L_1}X_{L_2L_2} + X_{L_2}X_{L_1L_1}).$$

Since the assumption of production functions that are strictly quasi-concave and homogeneous of degree 1 implies that $X_{L_jL_j} < 0$, we know that $(d^2X_2/dX_1^2)^{SR} < 0$.

change in commodity prices along a given SR transformation curve:

$$\frac{dw}{w} = -\frac{X_{L_1L_1}}{\Delta}\frac{dp}{p}, \tag{10}$$

$$\frac{dr_1}{r_1} = \frac{X_{L_1L_1}}{\Delta}\left(\frac{wL_1}{r_1K_1}\right)\frac{dp}{p}, \tag{11}$$

$$\frac{dr_2}{r_2} = \left[1 - \frac{pX_{L_2L_2}}{\Delta}\left(\frac{wL_2}{r_2K_2}\right)\right]\frac{dp}{p}, \tag{12}$$

where we utilized the relations

$$X_{L_jK_j} = -k_jX_{K_jK_j} = -X_{L_jL_j}/k_j = X_{K_jL_j},$$[7]

and where

$$\Delta = -(X_{L_1L_1} + pX_{L_2L_2}) > 0 \quad \text{and} \quad k_j = K_j/L_j.$$

Our results are stated in percentage terms in order to indicate not only directions, but also relative magnitudes of factor-return changes.

From (10), it can be seen that the common wage rate, $w = w_1 = w_2$, is directly related to p. As long as relative commodity prices remain unchanged, the wage rate (expressed in terms of commodity 1) will not change either. On the other hand, if p rises, the wage rate will rise as well, although at a slower rate than the relative commodity price; that is, if $dp/p > 0$, then $dp/p > dw/w > 0$. As far as returns on capital are concerned, which in the SR may not be the same in both industries, (11) and (12) reveal that as the price of the second relative to the first commodity rises, the return on capital in the second industry increases at a faster rate than p, whereas the return on capital in the first industry declines; that is, if $dp/p > 0$, then $dr_2/r_2 > dp/p > 0 > dr_1/r_1$.

In summary, a tariff structure which leads to a *ceteris paribus* increase in the relative price of a given commodity will, in the SR, raise returns on capital and labor employed in this, the protected industry. In the unprotected industry, only wages will rise, while the capital return will fall. Clearly, there exists an incentive for a given industry to ask for tariff protection relative to other industries, since all factors employed in this industry will be able to raise their earnings.[8] And the more immobile factors employed, the greater the benefits the protected industry can enjoy, and the stronger the incentive for demanding tariff protection.

---

[7] These relations are implied by the assumption of homogeneity of degree 1 of the production functions.

[8] Since $dp/p > dw/w > 0$, income of workers in terms of commodity 1 will definitely rise. This, however, does not imply that workers are necessarily going to be better off than before. But the smaller the fraction of X_2 in the consumption basket of workers, the more likely it is that they are going to gain from the mentioned tariff change.

Our conclusions stand in sharp contrast to those of a LR impact of price changes on factor returns, namely, the Stolper-Samuelson theorem. In the LR, an increase in the relative price of a given commodity will lead to a more than proportionate increase in $w(r)$ and to a decrease in $r(w)$ if this commodity is labor (capital) intensive. Hence, one factor benefits from the price change, while the other factor definitely loses. It is not possible that in the LR all factors employed in a given industry will gain from protection. Consequently, conventional trade theory finds it difficult to explain why lobbyists of a given industry would push for tariff protection and be supported by workers and capital owners.

The LR equilibrium, as specified in (2) and (3), will be reestablished when enough capital has moved from the low to the high capital return industry such as to equate the value of the marginal product of capital in the two productive activities. This change in capital allocation, which in figure 1 amounts to a gradual downward shift of the SR transformation curve from A to B, is denoted by $d\bar{K}_1 = -d\bar{K}_2 = dS$; $dS \gtreqless 0$ depending on $r_1 \gtreqless r_2$. If $(1/p)$ rises, than at the new SR equilibrium point, a^B, r_1 exceeds r_2, and capital owners will find it profitable to transfer capital goods from the second to the first industry; the adjustment process to LR equilibrium has started. The adjustment path itself consists of the locus of SR equilibria, where each SR equilibrium point corresponds to a different capital allocation. The impact of the capital-stock adjustment process on the capital return in industry 1 can be traced out by going back to (9), differentiating the four equations totally while holding relative prices and labor endowments constant, and solving for:

$$dr_1/dS = X_{K_1K_1} + X_{K_1L_1}(X_{L_1K_1} + pX_{L_2K_2})/\Delta.$$

The first term on the right side measures the declining value of capital's marginal product in industry 1, whereas the second term represents the rising value of capital's marginal product as more labor is employed; the rise in labor employment is, of course, the consequence of the capital-allocation process.[9] Whether the total effect (dr_1/dS) is positive or negative depends on relative factor intensities, as can be seen after substituting the definition of Δ and again using $X_{L_jK_j} = -k_jX_{K_jK_j} = -X_{L_jL_j}/k_j = X_{K_jL_j}$:

$$dr_1/dS = -X_{K_1K_1}pX_{L_2L_2}\left(1 - \frac{k_1}{k_2}\right)/\Delta. \tag{13}$$

Equation (13) reveals that during the adjustment process, the return on capital in industry 1 will rise if its production process is relatively capital intensive and fall if it is relatively labor intensive; that is, $dr_1/dS \gtreqless 0$ as $k_1 \gtreqless k_2$.

---

[9] From (9), we also can derive that dL_1/dS, the optimal adjustment in industry one's labor use as capital is reallocated from the second to the first industry, equals

$$(X_{L_1K_1} + pX_{L_2K_2})/\Delta > 0.$$

Values for changes in r_2 and w during the capital-adjustment period can be obtained in a similar fashion from (9):

$$dr_2/dS = pX_{K_2K_2}X_{L_1L_1}\left(1 - \frac{k_2}{k_1}\right)/\Delta \gtreqless 0, \tag{14}$$

as $k_1 \gtreqless k_2$; and

$$dw/dS = -X_{L_1K_1}pX_{L_2K_2}(k_1 - k_2)/\Delta \lesseqgtr 0, \tag{15}$$

as $k_1 \gtreqless k_2$.

Equations (13)–(15) describe the adjustment of factor returns during the capital-reallocation process. In both industries capital returns rise and labor returns fall if the first industry is capital intensive.[10] Hence, the LR comparative-statics responses, as expressed in the Stolper-Samuelson theorem, come about through the gradual adjustment of the capital stock in the two respective industries, a process which leads to a rise (decline) in returns on the factor that is used relatively more (less) intensively in the production of the good whose relative price has increased.

Output responses to price changes have already been depicted in figure 1. An increase in $(1/p)$ leads to an immediate SR output effect, indicated by the movement from A to a^B. Only labor moves from the second to the first industry, whereas the inability of capital to reallocate instantaneously causes capital-return differentials. Slowly, as long as these differentials persist, capital will be transferred from the low- to the high-return industry, which in figure 1 amounts to a downward sliding of the SR along the LR transformation curve. The output adjustment path between LR equilibria is traced out by connecting all SR production points such as a^B and c^B, all of which are determined by the condition that the price line is tangent to a given SR transformation curve.

From the analysis of the previous paragraph, it follows that every LR output effect can be decomposed into an instantaneous SR and a time-absorbing adjustment effect. Stating in terms of elasticities, we can write:

$$\begin{aligned}\left[\frac{(1/p)dX_1}{X_1d(1/p)}\right]^{LR} &= \left[-\frac{pdX_1}{X_1dp}\right]^{LR} \\ &= \left[-\frac{pX_{L_1}}{X_1}\frac{dL_1}{dp}\right]^{SR} \\ &\quad - \left[\frac{pX_{L_1}}{X_1}\frac{dL_1}{dS} + \frac{pX_{K_1}}{X_1}\frac{dK_1}{dS}\right]\frac{dS}{dp}.\end{aligned} \tag{16}$$

---

[10] In LR equilibrium, $r_1 = r_2$. Since in the SR r_1 has risen faster than r_2, we also must show that during the adjustment period r_1 rises slower than r_2 if $k_1 > k_2$, and r_1 falls faster than r_2 if $k_1 < k_2$. A comparison of (13) with (14) and a few manipulations reveal that $dr_1/dS \lesseqgtr dr_2/dS$ depending on $(k_2 - k_1)/k_1 \lesseqgtr (k_2 - k_1)/k_2$.

The LR price elasticity of supply equals the SR plus the adjustment elasticity. Concerning the SR elasticity, we know that it must be positive since $dL_1/dp < 0$, as can be seen utilizing (9). With regard to the adjustment elasticity, the same positive sign must hold since $dS/dp < 0$, $dK_1/dS > 0$, and $dL_1/dS > 0$ as we know from footnote 8. Hence, the LR value of the elasticity of supply always exceeds the SR value as long as at least one factor is not perfectly mobile. *Ceteris paribus*, this implies that import elasticities of demand are smaller in SR than in the LR, a condition that, in turn, enhances the possibility of SR instability of trading systems.[11]

IV. Short-Run Responses to Changes in Labor Endowments and the Adjustment to Long-Run Equilibrium

Immigration laws, designed to enlarge the existing supply of labor, are in general opposed by workers and favored by capital owners. Workers fear that wages will decline and capital owners expect higher returns per unit of capital employed. Such reaction patterns stand in sharp contrast to predictions expounded by conventional trade theory, namely, that any change in factor endowments has no impact on factor returns at all, provided the economy remains nonspecialized. Again, it is the LR nature of standard trade theory that leads to this apparent contradiction. The divergence between actual anticipations of factor owners and predictions of theory disappear once we apply our SR trading model. In the SR, capital owners in both industries will gain from additional labor supply, whereas workers will lose.

In order to obtain the SR responses of factor returns to changes in labor endowment, we go back to (9), differentiate it totally, and solve for:

$$dr_1/dL = (pX_{L_1L_1}X_{L_2L_2})/(\Delta k_1) > 0, \tag{17}$$

$$dr_2/dL = (pX_{L_1L_1}X_{L_2L_2})/(\Delta k_2) > 0, \tag{18}$$

$$dw/dL = -(pX_{L_1L_1}X_{L_2L_2})/\Delta \quad < 0, \tag{19}$$

where we used the constant returns to scale property

$$X_{K_jL_j} = -X_{L_jL_j}/k_j$$

in order to simplify our results. In the SR, labor is reallocated instantaneously, whereas the capital allocation remains the same as at the initial LR equilibrium. Since both industries absorb more labor, capital returns must rise in both industries and the common wage rate must decline.

---

[11] Other modifications of the standard two-commodity, two-factor model, such as the introduction of intermediate goods, may lead to greater stability of the trading system; see Chang and Mayer (1973).

A comparison of (17) with (18), furthermore, reveals that the SR rise in returns on capital is greater in the labor- than in the capital-intensive industry; that is, $dr_1/dL \gtreqless dr_2/dL$, depending on whether $1/k_1 \gtreqless 1/k_2$.

The fact that in the SR capital returns rise in both industries provides a plausible explanation of why capital owners would favor liberalized immigration laws. The fact that SR capital returns in the two industries rise at different rates, on the other hand, provides the starting point for adjustment to LR equilibrium. It creates an incentive to transfer capital from the capital- to the labor-intensive industry; or $dS \gtreqless 0$ as $(1/k_1) \gtreqless (1/k_2)$. During the capital-reallocation period, factor returns again are changing in accordance with the process described in (13)–(15). Capital returns in both industries are declining and the common wage rate is rising.[12] Hence, the SR impact on factor returns, a rise in the return on capital and a fall in the wage rate, is counteracted by the adjustment impact, a decline in capital returns and a rise in the wage rate.[13] Finally, in LR equilibrium, as stated in (2) and (3), returns on capital and labor are again what they were before the inflow of labor occurred.

The total amount of capital that must be transferred before LR equilibrium is attained can be determined by going back to the labor constraint (3) and rewriting it as

$$K_1/k_1 + K_2/k_2 = L. \tag{3'}$$

As we remember that in the LR wage-rental ratios and, therefore, capital-labor ratios of the two industries are not affected by endowment changes, we take (3′), differentiate it totally, and set $dk_j = 0$:

$$dK_1/k_1 + dK_2/k_2 = dL.$$

After we substitute $dK_1 = -dK_2 = dS$, we solve for:

$$dS = [k_1k_2/(k_2 - k_1)]\, dL, \tag{20}$$

which relates the magnitude of capital to be reallocated to the initial change in labor endowment. *Ceteris paribus*, the more similar factor intensities in the two respective industries, the greater the need for extensive capital-stock reallocation.

With respect to changes in output of final commodities, the SR effects differ from the LR effects in a very substantive manner as well. From LR trade theory we know that output of the labor-intensive commodity rises at a faster rate than labor supply, while output of the capital-intensive commodity declines.[14] Hence, in the LR one industry expands and the other contracts as more labor is made available to the economy. In the SR, however, the result will be expansion of both industries, as we are going to show next. First, we go back to the industry production functions, stated in (1), and differentiate them with respect to L, allowing for changes in labor use only:

---

[12] If, for example, k_2 exceeds k_1, the resulting capital adjustment is measured by $dS > 0$. And since, for $k_2 > k_1$, $dr_j/dS < 0$, and $dw/dS > 0$ from (13) to (15), it is implied that both r_j decline and w rises.

[13] By comparing (13) with (14), it can be seen that during the adjustment period returns on capital decrease faster in the labor-than in the capital-intensive industry.

$$dX_j/dL = X_{L_j}(dL_j/dL), \qquad j = 1, 2. \tag{21}$$

The SR change in labor employment in industry 1 can be obtained by differentiating (9) and solving for:

$$(dL_1/dL) = -(pX_{L_2L_2})/\Delta > 0. \tag{22}$$

The SR change in employment of labor in industry 2 is derived from $(dL_2/dL) = 1 - dL_1/dL$, into which we substitute (22),

$$(dL_2/dL) = -X_{L_1L_1}/\Delta > 0. \tag{23}$$

Combining (22) and (23) with (21) yields:

$$dX_1/dL = -X_{L_1}pX_{L_2L_2}/\Delta > 0, \tag{24}$$

$$dX_2/dL = -X_{L_2}X_{L_1L_1}/\Delta > 0, \tag{25}$$

which shows that in the SR an increase in the supply of labor leads to expansion of both industries, irrespective of relative factor intensities.

Finally, we decompose the LR output effect, caused by a change in labor endowment, into a SR and an adjustment effect:

$$\begin{aligned}(dX_j/dL)^{\mathrm{LR}} = {} & X_{L_j}(dL_j/dL)^{\mathrm{SR}} \\ & + [X_{L_j}(dL_j/dS) + X_{K_j}(dK_j/dS)]\,\frac{dS}{dL}.\end{aligned} \tag{26}$$

The first term on the right side, the SR effect, must always be positive as we know from (24) and (25). The second term, the adjustment effect, can have either sign, depending on relative factor intensities. Inspection reveals that the adjustment effect is positive if X_j is relatively labor intensive and negative if X_j is relatively capital intensive. For example, if we set $j = 1$ and then substitute

$$dL_1/dS = -(X_{L_1L_1}/k_1 + pX_{L_2L_2}/k_2)/\Delta$$

from footnote 8, $dS/dL = k_1k_2/(k_2 - k_1)$ from (20), and $dK_1/dS = 1$, the adjustment effect of output for commodity one can be described by:

$$\begin{aligned}-(X_{L_1}X_{L_1L_1}/k_1 + X_{L_1}pX_{L_2L_2}/k_2 \\ + X_{K_1}X_{L_1L_1} + X_{K_1}pX_{L_2L_2})k_1k_2/(k_2 - k_1)\ \Delta.\end{aligned}$$

---

[14] See Rybczynski (1955) or Jones (1965).

Clearly, this expression is positive if $k_2 > k_1$ and negative if $k_2 < k_1$. Hence, the adjustment process leads to expansion of the labor-intensive industry over and above the SR expansion, whereas the capital-intensive industry will contract to offset the initial SR expansionary impact.

As we now continue with our evaluation of the LR output effect for the first industry, we can substitute (24) and the above-obtained expression for the adjustment effect into (26), simplify the terms, and state:

$$\begin{aligned}(dX_1/dL)^{LR} &= k_2[(L_1X_{L_1})/L_1 + (K_1X_{K_1})/L_1]/[k_2 - k_1], \\ &= k_2(X_1/L_1)/(k_2 - k_1),\end{aligned} \tag{27}$$

which is positive if the first industry is labor intensive and negative if the first industry is capital intensive.[15]

References

Atsumi, H. "The Long-Run Offer Function and a Dynamic Theory of International Trade." *J. Internat. Econ.* 1, no. 3 (August 1971): 267–300.

Chang, W., and Mayer, W. "Intermediate Goods in a General EquilibriumTrade Model." *Internat. Econ. Rev.* 14, no. 2 (June 1973): 447–59.

Haberler, G. *The Theory of International Trade, with Its Applications to Commercial Policy.* Translated by A. Stonier and F. Benham. London: Hodge, 1936.

Jones, R. W. "The Structure of Simple General Equilibrium Models." *J.P.E.* 73 (December 1965): 557–72.

———. "A Three-Factor Model in Theory, Trade, and History." In *Trade, Balance of Payments and Growth: Papers in International Economics in Honor of Charles P. Kindleberger*, edited by J. N. Bhagwati, R. W. Jones, R. A. Mundell, and J. Vanek. Amsterdam: North-Holland, 1971.

Kemp, M. C. *The Pure Theory of International Trade and Investment.* Englewood Cliffs, N. J.: Prentice-Hall, 1969.

Lancaster, K. *Mathematical Economics.* New York: Macmillan, 1968.

Rybczynski, T. N. "Factor Endowments and Relative Commodity Prices." *Economica*, n.s. 22 (November 1955): 336–41.

[15] This is precisely the expression given in Kemp's derivation of the output effect in response to endowment changes (1969, p. 14).

Stability of the Demand for Money during the Great Contraction—1929–1933

Arthur E. Gandolfi

First National City Bank

This paper uses cross-sectional state data to test the stability of the demand for money (total bank deposits) during the Great Contraction. The demand for deposits is defined as a function of state permanent income, the rate of interest paid on deposits, and the rate of state bank failures. Separate yearly demand functions are estimated for each year of the period, and these results strongly suggest that the demand for deposits was stable over this period and that the contraction had a negligible effect on the size and significance of the income and interest-rate elasticities.

I. Introduction

The failure of the Federal Reserve System to stem and reverse the course of the 1929–33 contraction was widely interpreted at that time to mean that monetary policy was an instrument of limited usefulness and was inherently incapable of dealing with such periods of severe unemployment. Velocity was believed to be either extremely unstable or essentially indeterminate, responding passively to changes in money and income. While the belief that increases in the stock of money would have little or no expansionary effect on income during depressions was not limited to any particular "school" of economics, it found its most famous and detailed expression in the Keynesian notion of the liquidity trap (absolute liquidity preference). According to this theory, the functional relationship between money and income is destroyed during certain periods because

This paper is in part based on my Columbia University Ph.D. dissertation, "The Demand for Commercial Bank Deposits during the Great Contraction." I am grateful to Donald Dewey and Phillip Cagan for their many helpful comments and suggestions. I would also like to thank my colleagues at First National City Bank, especially Dudley D. Johnson and James R. Lothian, for their valuable substantive and stylistic comments.
[*Journal of Political Economy*, 1974, vol. 82, no. 5]
© 1974 by The University of Chicago. All rights reserved.

the public becomes willing, at a given level of income, to hold any amount of money supplied at a constant rate of interest.

Economists associated with the monetarist or "quantity theory school" have argued that the ineffectiveness of monetary policy during the Great Depression was in fact the result of the Federal Reserve's failure to use its powers to reverse or even halt the contraction that was occurring in the supply of money rather than the consequence of any breakdown or instability in the functional relationship between money and income.

The purpose of this paper is to investigate further these two competing hypothesis for the period 1929–33. All other research into the demand for money during the Great Depression has involved the use of aggregate time-series data. By contrast, this paper, like Edgar Feige's (1964) study of the postwar period, utilizes cross-sectional state data. We estimate yearly demand functions for money for the years 1929–33 and use these estimates to test the stability of the demand for money and the possible existence of a liquidity trap in the period under consideration.

II. The Demand-for-Money Function

Since data on currency holdings are not available for individual states, total adjusted deposits at commercial banks were used as a proxy for money.[1] As a model of the demand for money we assumed the following loglinear relationship:

$$\log \frac{(D)}{(P)} = \log \beta_0 + \beta_1 \log \frac{(Y^p)}{(P)} + \beta_2 \log C^d, \qquad (1)$$

where D is per capita total bank deposits, Y^p is permanent per capita nominal personal income, C^d is one plus the opportunity cost of holding bank deposits, and P is the permanent price-level variable.[2] This form was chosen because of its widespread use in other studies, because it gives direct estimates of the elasticities of deposits with respect to the independent variables, and because the logarithmic transformation can be expected to make the data for the different states more nearly homoscedastic.

---

[1] Total adjusted bank deposits include demand and time deposits corrected for float, less interbank and federal government deposits. State deposit holdings for any year are estimated by the amount of deposits held as of June 30 of that year. We chose total deposits in preference to demand deposits alone because during this period the meaning of demand and time deposits varied with the legal category of the bank (member or nonmember of the Federal Reserve System) and in the case of nonmember banks, with the state; see Friedman and Schwartz (1970).

[2] In calculating the permanent income series, we used an exponentially declining weighting system that, like Milton Friedman's, has an initial weight of .33. Appendix A describes the construction of this series in detail and contains a more complete discussion of the data and their sources.

The C^d is defined as the ratio of the net value (in the next period) of a dollar invested in the national asset market to the net value (in the next period) of a dollar held as deposits, or

$$C^d = \frac{(1 + R^n)}{(1 + R^d)(1 - L^e)}, \tag{2}$$

where R^n is the rate of interest paid on national financial assets, R^d is the rate paid on deposits, and L^e is the present value of the expected loss rate on deposits from bank failures and suspensions.[3] It should be noted that if the values of R^n, R^d, and L^e are sufficiently small, the logarithm of C^d is approximately equal to

$$R^n - (R^d - L^e), \tag{3}$$

the absolute difference between the net return on national assets and bank deposits.

The variable L^e is not equivalent to the expected failure rate, since not all the deposits in failed banks were necessarily lost forever; a large fraction of the deposits of failed banks were only frozen until the failed bank could be reopened or until its assets could be liquidated. To estimate L^e we assumed that the expected loss rate in any given state at time t is a function of the current and lagged bank-failure rates in that state. In particular we assumed

$$\text{Log}\,(1 - L_t^e) = \sum_{i=0}^{n} wi\,\text{Log}\,(1 - F_{t-i}), \tag{4}$$

where F_t is the percentage of deposits that failed in a state in year t.[4]

The inclusion of lagged failure-rate terms in the empirical demand function may also serve another purpose. A bank failure in a particular state will introduce a discrepancy between actual and desired deposit holdings. While the depositors of that state can eliminate this difference by drawing reserves from the country at large, it seems unreasonable to suppose that equilibrium will necessarily be reached within a year. It is more likely that failed deposits will be restored only slowly over time. Since the expected loss-rate variable was defined as a function of current and past failure rates, the coefficients of these variables will capture both the expectation effect and the direct disequilibrium effect of bank failures on the demand for deposits.[5]

---

[3] This formulation of the opportunity cost of holding deposits assumes that all differences in expected rates of price change across states are reflected in R^d and ignores any relationships that might exist between deposits and the liabilities of nonbank financial intermediaries. Since actual rates of price change differ little from state to state, the first assumption seems to be a valid simplification. The second is merely an empirical expedient—data on rates paid on such assets as the deposits of mutual savings banks and savings-and-loan shares are not available for this period.

[4] To the extent that failures in one state influence deposits holdings in the nation as a whole, the effect of failures on the demand for deposits will be captured by the constant terms of the yearly cross-section regressions.

The formulation of the demand function presented in equation (1) and those used by most demand-for-money studies assume that the appropriate variables are real per capita deposit (money) holdings and real per capita income. This structure is based on the conviction that the demand for money is homogeneous to the first degree with respect to prices (absence-of-money illusion), but not necessarily with respect to real income.

Unfortunately, no data are available on state price levels, and we have to use nominal data for deposits and income. If state price levels are identical, this procedure is valid, and the separate effect of prices will be captured by the constant term in each of the yearly equations. However, if prices are not constant across states, we will not necessarily obtain unbiased estimates of the elasticities of real deposits with respect to real income.

The degree of bias depends both upon the true value of the real permenent income coefficient and the relationship between real permanent state income and state price levels.[6] More specifically,

$$E(\hat{\beta}_1) = \beta_1 + (1 - \beta_1)\frac{\lambda_1}{1 + \lambda_1}, \tag{5}$$

where $\hat{\beta}_1$ is the income coefficient estimated from nominal permanent income and deposit data, β_1 is the true value of the elasticity of real deposits with respect to real permanent income, and λ_1 is the elasticity of state permanent prices with respect to state real permanent income and is assumed to be nonnegative. The degree to which $\hat{\beta}_1$ differs from β_1, therefore, depends both on the size of λ_1 and on the amount of the divergence between β_1 and unity. Any test of the stability of β_1, using $\hat{\beta}_1$, will therefore depend upon the stability of λ_1 over time.

III. Empirical Results

Table 1 presents the results of estimating the demand-for-deposit function defined in equation (6) for each of the 5 years of the 1929–33 contraction:

$$\begin{aligned}\text{Log}\, D_t = \beta_0 + \beta_1 \text{Log}\, Y_t^p + \beta_2 \text{Log}\,(1 + R_t^d) + \beta_3 \text{Log}\,(1 - F_t) \\ + \beta_4 \text{Log}\,(1 - F_{t-1}) + \beta_5 \text{Log}\,(1 - F_{t-2}) + \varepsilon.\end{aligned} \tag{6}$$

[7]

---

[5] The lagged failure-rate terms are defined as the log of one minus the state deposit failure rate. A coefficient of zero for a failure-rate term means that all failed deposits have been restored since the failures occurred, while a coefficient of unity implies that none has. If a coefficient exceeds unity, the excess can unambiguously be attributed to the reduction in deposit holdings caused by increased expectations of future failures.

[6] For the derivation of the relationship, see Appendix B.

[7] Since this is a cross-sectional equation, the rate of interest on national assets has not been included. Its effects and the effects of all national variables will be captured by the constant term.

TABLE 1

YEARLY DEMAND FUNCTIONS FOR TOTAL DEPOSITS (1929–33)

	Coefficient	S.E.	*t*-Value	
1929:				
Constant	−3.073	0.480	−6.403	
Log Y_t^p	1.262	0.074	17.092	$\bar{R}^2 = .900$
Log $(1 + R_t^d)$	26.995	6.785	3.978	$R^2 = .911$
Log $(1 - F_t)$	2.988	2.463	1.213	S.E. = .1651
Log $(1 - F_{t-1}$	−4.853	2.995	−1.620	
Log $(1 - F_{t-2})$	−0.384	1.569	−0.245	
1930:				
Constant	−3.123	0.404	−7.727	
Log Y_t^p	1.260	0.061	20.522	$\bar{R}^2 = .921$
Log $(1 + R_t^d)$	30.406	6.123	4.966	$R^2 = .929$
Log $(1 - F_t)$	1.808	0.788	2.296	S.E. = .1517
Log $(1 - F_{t-1})$	2.642	2.164	1.221	
Log $(1 - F_{t-2})$	−6.102	3.131	−1.949	
1931:				
Constant	−3.493	0.542	−6.450	
Log Y_t^p	1.312	0.082	16.039	$\bar{R}^2 = .914$
Log $(1 + R_t^d)$	31.641	7.065	4.479	$R^2 = .923$
Log $(1 - F_t)$	0.521	0.512	1.017	S.E. = .1728
Log $(1 - F_{t-1})$	1.029	0.753	1.367	
Log $(1 - F_{t-2})$	−0.626	1.955	−0.320	
1932:				
Constant	−3.428	0.561	−6.115	
Log Y_t^p	1.299	0.085	15.353	$\bar{R}^2 = .907$
Log $(1 + R_t^d)$	33.289	8.095	4.113	$R^2 = .917$
Log $(1 - F_t)$	1.655	0.537	3.082	S E. = .1915
Log $(1 - F_{t-1})$	0.293	0.563	0.521	
Log $(1 - F_{t-2})$	0.455	0.723	0.630	
1933:				
Constant	−3.013	0.578	−5.217	
Log Y_t^p	1.284	0.089	14.417	$\bar{R}^2 = .898$
Log $(1 + R_t^d)$	21.316	9.477	2.249	$R^2 = .908$
Log $(1 - F_t)$	0.799	0.147	5.433	S.E. = .2065
Log $(1 - F_{t-1})$	1.696	0.588	2 884	
Log $(1 - F_{t-2})$	0.745	0.612	0.122	

The coefficients of the permanent-income variables are the elasticities of deposits with respect to permanent income. These elasticities are significantly greater than unity at the 95 percent confidence level for every year of the period under consideration.[8] More important, they fall within the narrow range of 1.26 (for 1930) to 1.31 (for 1931).

In addition to the obvious stability of the estimated-income coefficients over this period, the *t*-values of these estimates remain very high. There is no evidence of any substantial change in the relationship between income and deposit holdings due to the economic contraction. If a liquidity trap had in fact existed during 1 or more years of this period, there would have been substantial differences in the size and significance of the estimated-income coefficients from 1 year to the next. In a liquidity trap, state deposit holdings are determined by supply, not demand, conditions, and there is no reason to believe that income would affect the supply of state deposits in the same way it affects their demand.

---

[8] Since the estimated values of the income coefficients exceed unity, the use of nominal instead of real income and deposit data tends, if anything, to bias the estimated values of the income coefficients downward. However, a casual examination of state data on income and price changes suggests that little relationship exists between these two variables, and λ_1 should be close to zero.

The coefficients of the interest-rate variable are significant and positive as expected, since an increase in the rate of interest paid on bank deposits reduces the opportunity cost of holding deposits and thus increases the amount held. The behavior of interest-rate coefficients during this period provides further evidence against the existence of a liquidity trap. Examination of the estimated interest-rate coefficients in table 1 shows that these did not substantially increase in size over the 1929–33 period, as would happen if the economy were in, or moving toward, a liquidity trap.

The interest elasticities implied by the coefficients estimated in table 1 tell the same story. The interest elasticities (estimated at the means) for the years 1929–33 are .54, .61, .65, .62, and .36, respectively. These obviously do not exhibit any large upward movement as the liquidity-trap hypothesis would imply.[9] The coefficients of the failure-rate variables do not demonstrate any obvious stability. Unlike the coefficients for the income and interest-rate variables, the coefficients of the failure-rate terms are rarely significant, often possess the wrong sign, and change in value from year to year. But this is not surprising, since there is a great deal of collinearity among the lagged failure-rate terms in each equation and since the nature of bank failures changed considerably over the period under consideration. For example, in the early part of the period, most failures occurred in small rural banks, and it was not until after June 1931 that large urban banks failed in any great numbers.

IV. Testing the Stability of the Demand-for-Deposits Function

One advantage of estimating a demand function using cross-sectional data is that many variables that vary over time are constant for all observations at any given time. Because of this, we did not include the national interest rate, the price level, the expected rate of price change, and other such variables in our estimating equation. Changes in these "national" variables, over time, will cause corresponding changes in the constant term in each of our yearly regressions. Thus, a function estimated using cross-sectional data may have significantly different constant terms for different years even if the function, defined to include the national variables, is completely stable. Since we jointly test for the stability, over time, of the intercepts and the coefficients of the variables included in the individual equations, we are in fact implicitly testing whether or not changes in "national" variables had a significant effect on each state's demand for deposits during this period.

---

[9] Our failure to find any evidence that would suggest the existence of a liquidity trap in the early thirties using cross-sectional data supports similar conclusions reached by Meltzer (1963) using annual time-series data. He found no empirical support for the liquidity-trap hypothesis that the interest elasticity of money should increase as the level of long-term interest rates declines.

To test the temporal stability of the money-demand function, we pooled all the observations for the 5-year period, estimated one demand equation, and then compared the sum of squared errors of that equation with the sums of squared errors from the separately estimated yearly equations.[10] The equation estimated from the pooled data was the following:

$$\begin{aligned} \text{Log } D_t = \; & \underset{(-15.0)}{-3.193} + \underset{(39.2)}{1.276} \text{ Log } Y_t^p + \underset{(9.1)}{29.01} \text{ Log } (1 + R_t^d) \\ & + \underset{(7.1)}{.741} \text{ Log } (1 - F_t) + \underset{(3.2)}{.927} \text{ Log } (1 - F_{t-1}) \\ & \underset{(-\,.13)}{-\,.047} (1 - F_{t-2}). \end{aligned} \tag{7}$$

$$\bar{R}^2 = .919$$

$$R^2 = .920$$

$$\text{S.E.} = .18012 \text{ (}t\text{-values in parentheses)}$$

All of the variables in this equation, except for the last of the failure-rate variables, are of the expected sign and are significant at the 95 percent confidence level. More important, these coefficients are not statistically different from those we estimated for the yearly equations (table 1). A test of the null hypothesis, that the demand for money function was stable over the 5-year period, yields an F-ratio of 1.114 with 24 and 215 df. This implies that we cannot reject the null hypothesis at either a 90 or 95 percent level of confidence.[11] Hence we conclude that the demand function for total deposits, as we have formulated it, was stable during this period.

Even though we cannot reject the hypothesis that the entire demand function is stable over the 1929–33 period, it is possible that the coefficients of one or more variables changed over time. To test for this possibility, we reestimated our pooled equation, this time including dummy variables to allow the intercepts and the coefficients for all the independent variables to differ for each of the 5 years in the

---

[10] The test is described by Kullback and Rosenblatt (1957).

[11] Note that $F(24, \infty)$, equals 1.38 at a 90 percent confidence level and $F(24, \infty)$ equals 1.52 at a 95 percent level.

TABLE 2

POOLED DEMAND EQUATION ESTIMATED WITH DUMMY VARIABLES*

	Coefficient	S.E.	t-Value
Constant	−3.030	0.478	−6.338
D 30	−0.030	0.672	−0.044
D 31	−0.395	0.704	−0.561
D 32	−0.458	0.701	−0.654
D 33	−0.019	0.646	−0.030
Log Y^p	1.254	0.072	17.341
D 30 Log Y^p	−0.007	0.102	−0.072
D 31 Log Y^p	0.049	0.107	0.453
D 32 Log Y^p	0.051	0.106	0.475
D 33 Log Y^p	0.036	0.098	0.370
Log $(1 + R^d)$	27.264	7.235	3.769
D 30 Log $(1 + R^d)$	4.360	10.169	0.429
D 31 Log $(1 + R^d)$	4.254	10.250	0.415
D 32 Log $(1 + R^d)$	6.810	10.344	0.658
D 33 Log $(1 + R^d)$	−6.101	10.883	−0.561
Log $(1 - F_t)$	2.790	2.534	1.101
D 30 Log $(1 - F_t)$	−1.835	2.648	−0.693
D 31 Log $(1 - F_t)$	−2.242	2.587	−0.867
D 32 Log $(1 - F_t)$	−1.125	2.582	−0.436
D 33 Log $(1 - F_t)$	−1.992	2.537	−0.785
Log $(1 - F_{t-1})$	−5.043	3.060	−1.648
D 30 Log $(1 - F_{t-1})$	5.070	3.648	1.390
D 31 Log $(1 - F_{t-1})$	5.946	3.132	1.898
D 32 Log $(1 - F_{t-1})$	5.343	3.105	1.721
D 33 Log $(1 - F_{t-1})$	6.746	3.102	2.175

$\bar{R}^2 = .921$
$R^2 = .928$
S.E. = .1779

* Dummy variables allow constant terms and coefficients for the years 1930–33 to differ from those of the base year, 1929.

sample.[12] The result of estimating this equation is presented in table 2.[13]

The test resulting from this regression is for significant changes in coefficients relative to the coefficients estimated for a selected base year. The year 1929 was chosen as a base year for this test on the grounds that it is the most "normal" year of the period. If the dummy for a given variable in a given year is statistically different from zero, then that coefficient is significantly different from the corresponding coefficient in the 1929 equation.

The base-year coefficients estimated by this equation are, as expected, very similar to those estimated by the 1929 demand equation. In addition, of all the dummy variables that allow the coefficients of the independent variables to differ over time, only the coefficients of the second failure-rate term are close to being significant. The t-values of all the other dummies are below unity, and most are below .5.

---

[12] The independent variables (including the intercepts) for all but 1 of the years of the sample are each multiplied by a separate dummy variable that takes the value of unity during that year and zero during all others.

[13] Because of computer limitations, only two failure-rate variables were included in the equation.

As mentioned above, the stability of the constant term, over time, has a special significance for a demand function estimated with cross-sectional data. Since the dummy variables for the yearly constant terms are not significantly different from zero, we conclude that changes in national variables (e.g., interest rates) did not have a significant net effect on deposit holdings during this period. This result is to some degree disconcerting. We would expect that the fall in national prices and interest rates, everything else being equal, should increase the demand for deposits. One explanation of our results is that the positive effect of these national variables was approximately offset by a growing national concern for the safety of deposits that was not captured by each state's failure-rate variables.

In addition to directly testing the stability of the demand for deposits for the years 1929–33, we can compare our cross-sectional estimates with those from a later period to see if there was a substantial change in the demand function over this longer interval. As mentioned above, Feige used state data to estimate the demand for financial assets for the period 1949–59. While Feige also found the demand relationships stable, his study differed considerably from this one in that he estimated separate linear demand functions for time and demand deposits and was able to obtain state data on rates paid on alternative financial assets. Since Feige was primarily concerned with estimating interest elasticities, he included a large number of state and regional dummy variables in his demand functions.[14] Pooling his cross-sectional data for his entire period, Feige estimated permanent income elasticities for time and demand deposits of .69 and .92, respectively. These elasticities are considerably lower than our separate estimates of the income elasticities for time (1.34) and demand deposits (1.24) using our pooled data.[15]

Lee (1966) reestimated Feige's pooled equation without the dummy variables and estimated an income elasticity of time deposits equal to 1.24 and of demand deposits equal to 1.36. Furthermore, Feige (in press) reestimated his pooled demand-deposit equation in loglinear form without dummy variables for the years 1949–65 and obtained an income elasticity of 1.20. These estimates are sufficiently close to our own that we are unable to find any substantial difference in the deposit-demand functions estimated in these two distinct periods.

[14] If we include dummy variables for New York and Illinois in our total deposit equations, the income coefficients are reduced slightly. These coefficients range from 1.20 (1933) to 1.24 (1931) and are significantly greater than unity.

[15] Our time- and demand-deposit equations are also stable over the 1929–33 period, but the coefficients of the interest-rate variables are not significant and often possess the wrong sign.

TABLE 3

YEARLY DEMAND FUNCTIONS FOR TOTAL DEPOSITS ESTIMATED USING MEASURED INCOME (1929–33)

	Coefficient	*S.E.*	*t-Value*	
1929:				
Constant	−3.073	0.480	−6.403	$\bar{R}^2$ = .900 R^2 = .911 S.E. = .1651
Log Y_t	1.262	0.074	17.092	
Log $(1 + R_t^d)$	26.995	6.785	3.978	
Log $(1 - F_t)$	2.988	2.463	1.213	
Log $(1 - F_{t-1})$	−4.853	2.995	−1.620	
Log $(1 - F_{t-2})$	−0.384	1.569	−0.245	
1930:				
Constant	−2.326	0.373	−6.232	$\bar{R}^2$ = .919 R^2 = .927 S.E. = .1539
Log Y_t	1.153	0.057	20.204	
Log $(1 + R_t^d)$	30.543	6.210	4.919	
Log $(1 - F_t)$	2.137	0.800	2.672	
Log $(1 - F_{t-1})$	3.702	2.184	1.695	
Log $(1 - F_{t-2})$	−7.574	3.183	−2.380	
1931:				
Constant	−2.014	0.609	−3.309	$\bar{R}^2$ = .860 R^2 = .874 S.E. = .2207
Log Y_t	1.116	0.094	11.884	
Log $(1 + R_t^d)$	32.294	9.019	3.581	
Log $(1 - F_t)$	0.855	0.650	1.315	
Log $(1 - F_{t-1})$	1.287	0.963	1.336	
Log $(1 - F_{t-2})$	−1.092	2.528	−0.432	
1932:				
Constant	−1.940	0.530	−3.658	$\bar{R}^2$ = .886 R^2 = .898 S.E. = .2126
Log Y_t	1.138	0.084	13.533	
Log $(1 + R_t^d)$	30.306	8.994	3.370	
Log $(1 - F_t)$	1.533	0.599	2.560	
Log $(1 - F_{t-1})$	0.359	0.626	0.573	
Log $(1 - F_{t-2})$	0.947	0.799	1.186	
1933:				
Constant	−1.602	0.648	−2.473	$\bar{R}^2$ = .837 R^2 = .854 S.E. = .2605
Log Y_t	1.124	0.105	10.709	
Log $(1 + R_t^d)$	17.619	11.972	1.472	
Log $(1 - F_t)$	0.788	0.186	4.246	
Log $(1 - F_{t-1})$	1.836	0.741	2.476	
Log $(1 - F_{t-2})$	0.540	0.764	0.708	

V. Permanent and Measured Income

Friedman (1959) has suggested that cyclical fluctuations of velocity are primarily the result of the fact that the demand for money is actually a function of permanent income, while velocity is calculated using measured income. Our results of estimating deposit-demand functions for the severest economic contraction in American history support his view. Even though measured income is a good proxy for permanent income when state cross-sectional data are used, demand equations estimated using permanent income are superior in explanatory power to those estimated using measured income for the 4 years 1930–33 (see table 3 and table 1).[16] In addition to the superior explanatory power of such demand equations, these two tables show that over this period the elasticity of deposits with respect to permanent income is considerably more stable than the elasticity with respect to measured income. After 1929, the elasticity of deposits with respect to measured income falls sharply. This decline is exactly what would be predicted by an errors in variable model. The probability limit of the coefficient that estimates the elasticity of deposits with respect to measured income will, in such a model, be equal to:

$$\beta_1 \text{ var } (Y^p)/\text{var } (Y) + \beta_2 \text{ var } (Y^t)/\text{var } (Y),$$

where β_1 is the true coefficient of permanent income and β_2 is the true coefficient of transient income (Y^t). If the coefficient of transient income is smaller than that of permanent income, the coefficient of measured income would fall as the variance of transient income rose relative to the variance of measured income during the 1929–33 period.

Table 4 presents the results of including a measure of transient income in the yearly demand equations.[17] Since on a priori grounds we expect the relationship between transient income and deposit holdings to be nonnegative, we cannot reject, by use of a one-tail test, the null hypothesis that there is no relationship between deposit holdings and transient income. The coefficients of transient income are either small and insignificant or possess the wrong sign.

Similar results are produced when a transient income variable is included in the pooled demand equation estimated for the years 1930–33

$$\begin{aligned}
\text{Log } D_t = \; & \underset{(-21.29)}{-3.276} + \underset{(33.29)}{1.286} \text{ Log } Y_t^p + \underset{(7.95)}{29.10} \text{ Log } (1 + R_t^d) \\
& + \underset{(6.89)}{.756} \text{ Log } (1 - F_t) + \underset{(3.11)}{.969} \text{ Log } (1 - F_{t-1}) \\
& + \underset{(.27)}{.102} \text{ Log } (1 - F_{t-2}) - \underset{(-.77)}{.09} (\text{Log } Y_t - \text{Log } Y_t^p) \qquad (8)
\end{aligned}$$

$\bar{R}^2 = .919$

$R^2 = .920$

S.E. = .1830

---

[16] Since permanent income is defined by necessity to be equal to measured income in 1929, the demand functions estimated using permanent and measured income are identical for that year. Similarly, all regressions later estimated using transient income exclude data for that year.

[17] Transient income is defined for our purposes as the difference between the log of measured and the log of permanent income. This definition does not necessarily satisfy the assumption of the errors in variable model that permanent and transient income be uncorrelated.

TABLE 4

YEARLY DEMAND FUNCTION FOR TOTAL DEPOSITS ESTIMATED USING A TRANSIENT-INCOME VARIABLE (1930–33)

	Coefficient	S.E.	t-Value	
1930:				
Constant	−2.847	0.532	−5.350	
Log Y_t^p	1.224	0.077	15.987	
Log Y_t/Y_t^p	0.430	0.535	0.803	$\bar{R}^2$ = .920
Log $(1 + R_t^d)$	30.422	6.150	4.947	R^2 = .930
Log $(1 - F_t)$	1.933	0.806	2.397	S.E. = .1524
Log $(1 - F_{t-1})$	3.008	2.222	1.354	
Log $(1 - F_{t-2})$	−6.660	3.223	−2.067	
1931:				
Constant	−4.091	0.569	−7.195	
Log Y_t^p	1.381	0.083	16.749	
Log Y_t/Y_t^p	−0.791	0.325	−2.435	$\bar{R}^2$ = .923
Log $(1 + R_t^d)$	31.597	6.692	4.722	R^2 = .932
Log $(1 - F_t)$	0.532	0.485	1.096	S.E. = .1637
Log $(1 - F_{t-1})$	0.800	0.719	1.113	
Log $(1 - F_{t-2})$	0.288	1.889	0.153	
1932:				
Constant	−3.726	0.737	−5.059	
Log Y_t^p	1.327	0.097	13.752	
Log Y_t/Y_t^p	−0.285	0.451	−0.632	$\bar{R}^2$ = .906
Log $(1 + R_t^d)$	34.058	8.242	4.132	R^2 = .918
Log $(1 - F_t)$	1.703	0.546	3.119	S.E. = .1928
Log $(1 - F_{t-1})$	0.316	0.568	0.556	
Log $(1 - F_{t-2})$	0.345	0.749	0.461	
1933:				
Constant	−3.325	0.595	−5.588	
Log Y_t^p	1.303	0.088	14.821	
Log Y_t/Y_t^p	−0.535	0.317	−1.689	$\bar{R}^2$ = .902
Log $(1 + R_t^d)$	23.305	9.352	2.492	R^2 = .914
Log $(1 - F_t)$	0.808	0.144	5.610	S.E. = .202
Log $(1 - F_{t-1})$	1.715	0.576	2.977	
Log $(1 - F_{t-2})$	0.059	0.599	0.100	

The coefficient of the transient-income term in the pooled demand equation is very small, insignificant, and also possesses the wrong sign.

These results suggest that, for the period under examination, the demand for deposits is a function of permanent income and that there is no stable or significant positive relationship between deposit holdings and transient income.

VI. Summary and Conclusions

The use of cross-sectional state data permitted us to estimate seaparate deposit demand equations for each year of the great contraction (1929–33). From these estimates we conclude that the demand-for-deposit function was stable and that there is no evidence to suggest that the values of income and interest-rate coefficients differed in any substantial or significant way during the contraction from their 1929 values. The constancy in the values of the income and interest-rate coefficients over time strongly suggests that we in fact estimate demand rather than supply relationships throughout the period and that these relationships remained essentially unaltered.

The evidence presented in this paper casts serious doubt on the argument that the relationships between money and income is so unstable (particularly during depressions) that monetary policy (changes in the supply of money) is of secondary importance. Our results show that there was no breakdown in the relationship between money and income from 1929 to 1933 and that the contraction did not destroy the foundation (a stable money-demand function) for an effective contracyclical monetary policy. In fact we cannot help but be impressed by how little the massive economic dislocations of the Great Depression seem to have affected the demand for money.

Appendix A

Data Sources

The data on bank deposits was obtained from *All Bank Statistics*, a work of the Federal Reserve System (1959). This source gives yearly estimates of the deposits of commercial banks in each state as of June 30. Though we are unable to make a calculation of the size of each state's float, we approximate it by using cash items in the process of collection by a state as a proxy for the amount of checks that the public of that state has written but that have not yet been deducted from its balances.

Income Data

Data for the personal income variable comes from the Commerce Department's personal income estimates by states.[18] We define permanent or expected income for year t as follows:

$$Y_t^P = Y_{t-1}^P + \beta(Y_t - Y_{t-1}^P),$$

where Y_t^P is the level of income expected in year t and Y_t is measured income in year t. The income series begins in 1929, and in order to calculate a permanent income series, we assumed that permanent income for the year 1929 equaled measured income for that year.[19] This assumption appears to be reasonable, when we consider that the enormous changes that took place in state income during the early thirties dwarf any discrepancy between measured and permanent income that might have existed in the immediate predepression period.

---

[18] Schwartz and Graham (1956).

[19] Our measured-income figures are yearly totals centered in June. Since our deposit data are defined as of the end of each June, to calculate a permanent-income series appropriate to the deposit data, we centered the income data in the previous December by taking a simple average of our June-centered data.

The Rate of Interest Paid on Deposits

For the period under study, it was possible to obtain interest data only for national banks.[20] While national banks do not constitute a perfectly representative sample of all commercial banks, they are well represented among both rural and urban banks. To calculate the rate of interest paid on demand and time deposits at national banks, we divided the amount of interest paid on each type of deposit in any given fiscal year by a weighted average of December and adjacent June deposit figures. The rate of interest on total deposits was calculated by taking a weighted average of the rate paid on time and demand deposits by national banks during the fiscal year. The weights were based on the distribution of the two types of deposits at all commercial banks during that year.

Failure-Rate Data

The failure-rate variable for any given year is defined as the amount of deposits in banks that failed or suspended in the previous 12 months, divided by the amount of deposits in existence 12 months earlier.[21] Data on the deposits of failed banks were provided by the annual reports of the Comptroller of the Currency (1929–34). The deposits of failed banks, as of June 1933, include the deposits of all banks that failed prior to that June, plus the deposits of all unlicensed banks that failed to reopen by June 30, 1933.

Appendix B

The demand function we would like to estimate includes a state price-level variable:

$$\begin{aligned}\text{Log } D_t &= \beta_0 + \beta_1 \text{ Log } Y_t^p + (1 - \beta_1) \text{ Log } P_t + \beta_2 \text{ Log } (1 + R_t^d) \\ &\quad + \beta_3 \text{ Log } (1 - F_t) + \beta_4 \text{ Log } (1 - F_{t-1}) + \beta_5 \text{ Log } (1 - F_{t-2}) + U_t.\end{aligned}$$

Because of the absence of state price-level data, we in fact estimate the following equation:

$$\begin{aligned}\text{Log } D_t &= \beta_0 + \beta_1 \text{ Log } Y_t^p + \beta_2 \text{ Log } (1 + R_t^d) + \beta_3 \text{ Log } (1 - F_t) \\ &\quad + \beta_4 \text{ Log } (1 - F_{t-1}) + \beta_5 \text{ Log } (1 - F_{t-2}) + U_t^*,\end{aligned}$$

where $U^* = (1 - \beta_1)(\text{Log } P - \overline{\text{Log } P}) + U$.

In matrix notation where β is the vector of regression coefficients, X is the matrix of data observations on the independent variables, Y is the data vector for the dependent variable, U is a vector of the disturbance terms, and e is a vector of residuals:

$$\begin{aligned}Y &= X\beta + U^*, \\ Y &= X\hat{\beta} + e, \\ \hat{\beta} &= (X'X)^{-1}X'Y, \\ &= \beta + (X'X^{-1})X'U^*,\end{aligned}$$

---

[20] Comptroller of the Currency (1928–33).

[21] The bank failure-rate variables for the years 1927–29 were calculated by dividing the amount of bank failures, for each of these years, by the amount of deposits in existence in June of 1929.

and substituting for U^*

$$\hat{\beta} = \beta + (X'X)^{-1}X'U + (1 - \beta_1)(X'X)^{-1}X'(\text{Log } P - \overline{\text{Log } P}).$$

For the income coefficient, this implies that $E(\hat{\beta}_1) = \beta_1 + (1 - \beta_1)\alpha_1$, where α_1 is the expected value of the regression coefficient obtained by regressing the log of state permanent price levels on the log of nominal permanent income per capita.[22]

$$\text{Log } P_t = \alpha_0 + \alpha_1 \text{ Log } (Y_t^P) + U_t'.$$

Since by rearranging terms we can obtain

$$\text{Log } P_t = \frac{\alpha_0}{1 - \alpha_1} + \frac{\alpha_1}{1 - \alpha_1} \text{ Log } (Y_t^P/P_t) + \frac{U_t'}{1 - \alpha_1},$$

α_1 is equal to $\lambda_1/1 + \lambda_1$, where λ_1 is the true value of the elasticity of state permanent prices with respect to state real permanent income.

References

All Bank Statistics, United States, 1896–55. Washington: Board of Governors of the Federal Reserve System, 1959.

Comptroller of the Currency. *Annual Report.* Washington: Government Printing Office, 1929–34.

Feige, E. *The Demand for Liquid Assets: A Temporal Cross-Section Analysis.* Englewood Cliffs, N.J.: Prentice-Hall, 1964.

———. "Alternative Temporal Cross-Section Specifications of the Demand for Demand Deposits." *J. Finance* (in press).

Friedman, M. "The Demand for Money: Some Theoretical and Empirical Results." *J.P.E.* 67 (August 1959): 237–51.

Friedman, M., and Schwartz, A. *Monetary Statistics of the United States.* New York: Columbia Univ. Press, 1970.

Kullback, S., and Rosenblatt, H. M. "On the Analysis of Multiple Regression in *k* Categories." *Biometrica* 44 (June 1957): 67–83.

Lee, Tong Hun, "Substitutability of Non-Bank Intermediary Liabilities for Money: The Empirical Evidence." *J. Finance* 21, no. 3 (September 1966): 441–57.

Meltzer, A. "Demand for Money: Evidence from the Time Series." *J.P.E.* 71, no. 3 (June 1963): 219–46.

Schwartz, C., and Graham, R., Jr. *Personal Income by States since 1929.* Washington: Government Printing Office, 1956.

---

[22] Prices are assumed to be orthogonal with all the other independent variables.

The Screening Hypothesis and the Returns to Education

Richard Layard and George Psacharopoulos

London School of Economics

The screening hypothesis suggests that intereducational earnings differentials, even when standardized for differences due to noneducational factors, reflect no direct productivity-enhancing effects of education but only its effects as a device for signaling preexisting ability differences. But three predictions in the spirit of the hypothesis are not in fact borne out. First, rates of return to uncompleted courses (dropouts) are as high as to completed courses. Second, standardized educational differentials rise with age, although employers have better information about older employees' abilities. Third, if screening is the main function of education, it could be done more cheaply by simpler testing procedures. Yet these are not used widely despite the profits that could, according to the hypothesis, be made by those developing them.

The productive role of education has been questioned before, but is now under unusually heavy attack from the "screening hypothesis." If true, this hypothesis has quite devastating implications for educational policy and research. Broadly, it says that the earnings differentials associated with education do not mainly reflect improvements in individual productive capacity caused by education but, rather, employers' use of education to identify preexisting differences in talents. If education has any social value, it is as a signaling device which helps to place the right man in the right job. But even in this case, too much education is likely to be sought, with the private returns to education exceeding the social returns.

We are grateful to K. Arrow, M. Blaug, H. G. Johnson, W. McMahon, D. Metcalf, and P. J. Wiles for helpful comments on an earlier draft.
[*Journal of Political Economy*, 1974, vol. 82, no. 5]
© 1974 by The University of Chicago. All rights reserved.

The Hypothesis

The most rigorous version of the hypothesis is by Arrow (1973) who assumes for purposes of argument that individual productive ability is completely unaffected by education. Employers have no information about employees except their education, so education is used as the sole basis of pay. If the marginal product of graduates (treated as a class) is higher than that of nongraduates, due to their higher initial ability, graduates get paid more than nongraduates. So it pays someone with a good chance of graduation to go to college, as long as the private returns exceed the cost. We now distinguish two cases, one where different qualities of labor are perfect substitutes in production and the second where their productivity depends on what job they are in. In the first case, all education is socially valueless, since the productive contribution of the individual is not affected by what job he does and there is thus no return to improving employers' information about individual abilities. As a result, everybody could be better off if the more able would agree not to seek education in an attempt to signal their ability to employers. In the second case, education has social value up to a point by improving the allocation of people between jobs, but, depending on its cost, there may well be overinvestment such that no screening would be better for everyone than the amount provided at competitive equilibrium.

Other leading exponents of the screening hypothesis, such as Berg (1971), Taubman and Wales (1973), and Wiles (1974), put forward similar though less sharply delineated models, so we shall concentrate on Arrow's.[1] The problem is how to test it.

The Production Function Approach

If there were perfect substitution between different categories of labor (which there is not), the screening hypothesis could in principle be tested by production functions fitted across regions where labor forces differed only with respect to education. Interregional production functions which make this assumption have indeed been fitted by Griliches (1970) and others and show an effect of education on production similar to that indicated by wage differentials between people with different levels of education. However, within one country it is not correct to assume that the educational level of the labor force varies across states without any associated variation in noneducational labor force quality. More seriously, insofar as the labor force in one state is drawn disproportionately from those brought up in the same state, there is a simultaneity problem in that people who are richer may wish to "consume" more education even if it does not raise their output. This problem has not yet been fully tackled (but see Johnson [1970], Tolley and Olson [1971], and Fallon and Layard [in press]). Until more work has been done, one cannot claim production function results as very clear evidence against the screening hypothesis.

---

[1] See also Spence (1973) and Stiglitz (forthcoming). For another discussion of this issue in a different vein but a similar spirit, see Blaug (1972).

The Earnings Function Approach

One therefore turns naturally to earnings functions, where individuals provide the observations. Unfortunately, no simple answer can be found here either. If P is the individual's marginal contribution to output, S his schooling, and A a vector of ability and other relevant attributes not caused by education, we may suppose

$$P = f(S, A). \tag{1}$$

We wish to know f_S, the effect of schooling on productivity.

For purposes of an earnings function, however, research workers observe income Y and a different vector of ability variables A^R. They then estimate a function

$$Y = g(S, A^R) + u. \tag{2}$$

The question is whether f_S and g_S are the same.

Until recently, the main controversy concerned the problem of whether A^R sufficiently approximated A for g_S not to be biased upward owing to positive correlation between S and elements of A not included in A^R. The general finding of research, summarized in Appendix I, has been that the partial regression coefficient of earnings on education (g_S) is a high fraction of the simple regression coefficient where no attempt is made to control for ability.

But the screening hypothesis has added a quite new problem. Suppose schooling really has no effect on productivity ($f_S = 0$), but employers do not know the full vector A for their employees. Suppose also that the productivity of an individual cannot be directly measured, so that employers cannot pay a person his marginal product (P). They can, however, statistically predict it from a production relation

$$P = h(S, A^E) + v, \tag{3}$$

where A^E is the set of ability data available to employers.[2] If A^E is not the same as A, h_s may easily be positive even if f_S is zero. If the employer now paid an income (Y) equal to $\hat{P}$, the researcher's earnings function could show a positive effect of education ($g_S > 0$), even if he included in it the

---

[2] Alternatively, assume that, although individual productivity can be directly measured, hiring arrangements make it necessary to fix a person's wage before his productivity is known.

TABLE 1

ACTUAL AND FREE-ENTRY DISTRIBUTION OF EDUCATIONAL GROUPS BETWEEN OCCUPATIONS (MALES)

	PERCENTAGE		
EDUCATION	High-Pay Occupation*	Low-Pay Occupation†	Total
High school:			
Actual	61	39	100
Free entry	96	4	100
College:			
Actual	96	4	100
Free entry	98	2	100

SOURCE.—Taubman and Wales (1973), table 5.
* High-pay occupations are "professional, managerial, technical, and sales workers."
† Low-pay occupations are "blue collar, white collar, and service workers."

set of noneducational variables which exactly determine productivity ($A^R = A$).[3] Thus, education might wrongly appear to affect productivity whereas in fact it only affects earnings. *The issue is thus no longer whether education has an effect in a properly specified earnings function, but why it has an effect, if it does.*

Clearly, such an issue cannot be solved without data on A^E as well as A, A^R, and P. This is the fundamental problem with the pioneering study by Taubman and Wales (1973). Essentially they estimated function (2) for each occupational group, finding substantial education effects in each. They then hypothesized that in the absence of screening each individual would enter that occupation in which he could obtain the highest wage, given his schooling and other attributes. Hence, they produced a "free-entry" occupational distribution for each educational group, which is broadly compared with the actual distribution in table 1. As this shows, college-educated people achieved almost exactly the free-entry distribution, while those with only high school did far worse than under free entry. Taubman and Wales conclude that this was due to screening and that, under free entry, the private returns to education would have been one-half to two-thirds what they actually are. If screening is unproductive, the true social return to education is then a third to a half of that indicated by earnings differentials.

This ingenious analysis raises many problems. First, it is not clear what information employers are meant to have, except for S. We are told that, once in an occupation, people are paid their marginal product,[4] so there is presumably no screening problem of the kind described by Arrow and most other writers. This kind of problem only arises if an employer may end up paying a worker an amount different from his marginal product. Taubman and Wales assume that employers have no difficulty in paying people appropriately once they employ them. So screening presumably reflects not some objective difficulty of selection, but a taste for discrimination.

---

[3] Sufficient conditions for this are shown in Appendix 2.

[4] If people are paid their marginal product and screening is and will be practiced, then graduation does raise a person's productivity partly by pushing him through the screen, and this is a genuine social return (*pace* Taubman and Wales 1973, p. 50).

However, this leaves one crucial problem with their analysis. Why, if employers are willing to offer a given "high" wage to some high school graduates of a given quality, are they not willing to pay it to all high school graduates of that quality? It is surely not "screening" to discriminate between identical nongraduates who are, according to the model, able to produce the same output in the high-paying occupation. The truth of the matter is that the nongraduates who enter the top occupations are not the same (in terms of A variables, like motivation) as those with the same A^R who fail to enter them, and employers know this.

A quite different test of the screening hypothesis has been propounded by Berg (1971), who points to the rise over time in the educational qualifications found in each occupation. However, the nature of occupations changes over time, and such a test is no evidence of anything.

Three Unverified Predictions of the Screening Hypothesis

So what can one say, or should one say nothing? Unfortunately, the matter is of such importance that we have to operate with some provisional view of the significance of screening. No one, least of all Arrow, would maintain that all earnings differentials are in fact due to screening, so a test of his pure hypothesis, could it be devised, would not be very informative. Equally, no sane person would deny that the idea had some truth in it. What is needed is an estimate of the share of differentials due to it. Pending further empirical work, we shall examine three predictions which are in the spirit of the screening hypothesis and show that they are at variance with the evidence, including that of Taubman and Wales.

a) Private Returns Are to Certificates and Not to Years of Schooling

The screening hypothesis implies that some aspects of a person's educational record are more useful to an employer than others. In particular, graduation from a course should provide more evidence of ability and staying power than mere attendance for a number of years. This is the basis for the "sheepskin" version of the screening hypothesis according to which wages will rise faster with extra years of education when the extra year also conveys a certificate.

To test this, we compare the rates of return obtained by dropouts and by those who completed their courses.[5] The main available data are shown in table 2. In both the studies which allowed for the differential ability of dropouts, the rate of return to dropout exceeded that to completion. Taubman and Wales, though "surprised" at the finding, failed to note that it is in the least damaging to the screening hypothesis. In fact, it is devastating, unless the hypothesis says that employers use years of education rather than certificates as the screen.[6]

Two other studies explicitly claim to throw light on the role of certificates as compared with educational experience. The first concerns the effect of graduate study on earnings about 7 years after obtaining a bachelor's degree and uses a sample of 1,300 Woodrow Wilson fellows (Ashenfelter and Mooney 1968). The authors conclude that variables such as profession, degree level, and field of graduate study always explained more of the variance in earnings than years of graduate study and that this casts doubt on the "application of traditional rate of return analysis to the area of graduate education." However, the effect of graduate degree level (when compared with the effects of a bachelor's degree) consists of the positive effect of the Ph.D. (with t-statistic $= 2.5$) and the negative effects of the master's ($t = -0.9$) and of other higher degrees in law and the like ($t = -1.6$). Such strange effects can hardly be adduced in support of the screening hypothesis.

At the other end of the educational scale, Hansen, Weisbrod, and Scanlon (1970) claim to have cast doubt on the screening effect of high school graduation. They regress earnings of 2,400 low achievers rejected for the draft in 1963 on years of schooling, Armed Forces Qualification Test results, and age. They show that, holding AFQT and age constant, high school graduation does not significantly affect earnings. This, they claim, rejects the "sheepskin" hypothesis at the high school level, but the finding must be viewed with great caution for technical reasons pointed out by Chiswick (1972) and Masters and Ribich (1972).[7] Nevertheless,

---

[5] We would have preferred to show the earnings gain associated with each year of the course, including the year when it was successfully completed. Unfortunately, the data are not generally available in this form. They also generally show the rate of return to completion of one course over completion of the next lowest course rather than over noncompletion of the same course (see table 2).

[6] If the analysis is confined to students who are not "business owners," the rates of return to dropouts and graduates are the same. This is presumably because business owners have more earnings opportunities not directly dependent on their educational qualifications. The striking point is, however, the high rate of return to non-business-owning dropouts. In addition Taubman and Wales did an analysis which distinguished those who finished 1, 2, and 3 years of college. This showed a substantial earnings gain in the first year with none in the second or third, which might support a screening hypothesis which included college entry in the screen.

[7] The study also implies that education has virtually no effect (sheepskin or other) unless one is willing to accept Hansen, Weisbrod, and Scanlon's view that the correlation of AFQT and education arises mainly from an effect of education on AFQT, which seems unlikely.

TABLE 2

PRIVATE RATES OF RETURN TO DROPOUTS AND TO THOSE COMPLETING COURSES, MALES (UNITED STATES)

COURSE	RATE OF RETURN		SOURCE	SAMPLE	MAIN VARIABLES CONTROLLED FOR BY EARNINGS FUNCTIONS
	Dropout	Completed Course			
B.A.	15	11	Taubman and Wales (1973)	NBER sample of World War II veterans	Ability, father's education
B.A. + M.A.	8*	8			
High school	7	6	Rogers (1969)	Eighth graders in certain Connecticut schools, 1935	IQ, parental class, religion, marital status, number of jobs
B.A.	12	8			
High school	16.3	16.1	Hanoch (1967)	Census 1959, whites, North†	Place of residence, hours worked, marital status, family size, region of birth
B.A.	7.1	9.6			
High school	12.3	14.5	Hansen (1963)	Census 1949	None
B.A.	5.1	10.1			
B.A.	9.5	14.5	Becker (1964)	Census 1939, urban whites	None

---

* This figure is the rate of return on B A. + M A. courses for those who dropped out during the M A. course

† A further recent analysis of the 1959 census (Eckaus 1973) gives the following private rates of return for white males, North and South, adjusted for hours of work but not controlled for other variables by earnings functions

	Dropout	Completed Course
High school	3.5	4.0
B.A.	11.5	12.0

TABLE 3

STANDARDIZED EARNINGS OF THOSE WITH HIGHER EDUCATION AS A PERCENTAGE OF EARNINGS OF THOSE WITH COMPLETED HIGH SCHOOL: BY AGE (MALES)

	AVERAGE AGE		INCREASE IN DIFFERENTIAL
	33	47	
Some college	111	117	6
B.A.	112	131	19
Some graduate work	115	126	11
M.A.	110	132	22
Ph.D.	102	127	25
M.D.	172	206	34
LL.B	119	184	65

SOURCE.—Based on Taubman and Wales (1973, table 1). Data relate to NBER-TH sample of World War II veterans.

the opposite view—that certificates are crucial to the effect of education on earnings—is certainly not supported by the evidence.

b) Private Returns to Education Fall with Work Experience

A second prediction in the spirit of the screening hypothesis is that the partial effect of education on earnings, A^R held constant, will fall with experience as employers come to have better information about their employees' real productivity. (At the same time, the effect of A^R would rise.) Of course, one could proffer a screening hypothesis in which employers never learn, but this is not plausible. So we look to see whether the effect of education on earnings falls with experience, either absolutely or relatively.

If we ignore the problems caused by on-the-job training, we can at once reject the prediction of falling returns from Taubman and Wales's own findings, reproduced in table 3. This shows the earnings of those with each level of post-high school education, holding A^R constant, expressed as a percentage of average high school earnings, first when the sample was 33 and then when it was 47. As the table shows, the education effects rise both proportionately and absolutely with age. As one would expect, the proportional effect of measured ability also rises, but this does not reduce the effect of education, as the screening hypothesis would suggest. A similar pattern emerges from Hanoch's cross-sectional study based on census returns, reproduced in table 4, though in this case A^R is not held constant.

However, some would argue that the labor market is like a set of escalators. People are selected for a given escalator when they join the labor force and cannot thereafter easily jump from one escalator onto another. People with credentials are selected for escalators that rise rapidly and others for ones that move more slowly. People may of course walk at different speeds up their own escalator, but earnings

TABLE 4

STANDARDIZED EARNINGS: BY AGE (MALES, UNITED STATES, 1959)

Education	Age 22	27	37	47	57
Secondary (12 years) as % of elementary (8 years)	127	128	126	126	134
B.A. (16 years) as % of secondary (12 years)	...	126	144	161	161

SOURCE.—Hanoch (1967, p. 316)

differences between groups with different credentials are basically determined by the speeds at which their escalators are traveling.

However, the notion of rigidly separated escalators is not borne out by most studies of the relation of education and occupation, which reveal a wide variety of educational backgrounds among those in similar occupations (OECD 1969; Layard et al. 1971). Some occupations, like teaching, are of course more rigid than others; and there is unfortunately as yet little evidence on the correlation of starting salaries and lifetime incomes. However, our impression is that the correlation is far from perfect.

An alternative, and less institutional, explanation of why earnings differentials rise with age might attribute this to differential training (Mincer 1972). Suppose for example that training is employee financed and that graduates forgo a higher proportion of their potential starting salaries than nongraduates for this purpose. Then if workers were paid their marginal product, proportional earnings differentials between graduates and nongraduates would rise with experience. Now introduce screening so that earnings differentials are artificially raised in the early years of working life but not later. We might still find that proportional earnings differentials rise with experience, provided the difference in training is great enough. There is no conclusive answer to this argument, since the exact form of function relating training costs to experience is exceedingly difficult to determine empirically, even in the absence of the screening hypothesis. It is however up to the "screenists" to provide direct evidence of the importance of screening when there is prima facie evidence that casts doubt on it.

c) *Education Will Not Be Demanded If (Privately) Cheaper Screening Methods Exist*

We turn now from earnings data to consider the institutional implications of the screening hypothesis. According to the hypothesis, employers are willing to take on graduates at a substantial earnings differential because they consider this more profitable than employing nongraduates at a lower wage. This is simply because, although able nongraduates exist, graduates *are on average abler, for reasons not connected with their studies. The hypothesis therefore implies that the cost of discovering nongraduate* talent by suitably devised tests is at least as great as the observed discounted *earnings differential* between graduates and nongraduates *of the* same real ability.

Is this really plausible? One "screenist" (Wiles) believes it is not: "In all non-vocational cases a good one-day test of docility, perseverance and ability, costing about £20 instead of £2,000, would do as well [in revealing talent]." If this is so, why don't employers use it? Wiles's answer consists of two main points.

i) Educational Subsidies

First he draws attention to the public subsidy to higher education. However, this is irrelevant if, despite subsidies, employers still have to pay graduates a great deal more than nongraduates of the same measured ability—as the rates of return in table 2 indicate. True, measured ability is not the same as real ability, but casual observation confirms that employers do pay a good deal more for graduates than for nongraduates considered of equal natural ability. So why do they hire graduates if they could find equally able nongraduates at a lower wage? Wiles's answer is that the costs of finding able nongraduates, which could be small in a well-organized world, are in fact large.

ii) Unexploited Economies of Scale

He offers as an example a hypothetical table in which the total cost to employers of finding one able nongraduate is £600 as against £60 for a graduate. This difference occurs mainly because graduates have been pretested by the universities while nongraduates have not. But in this case, one must ask why no one private agency has established itself as a nongraduate testing body making its results available to all. The Educational Testing Service and American College Testing Program do this for academic abilities. If, as Wiles believes, similar tests can also pry out other abilities, there seems no reason why such a service has not developed —monopoly legislation is certainly not the reason why it has not.[8] We would not ourselves wish to decry the possibility of a disequilibrium in this market and the urgent need for government support for vocational guidance. But it is impossible to believe that a graduate differential as large as it is (holding ability constant) could be mainly explained by a failure to market some very cheap tests, plus a subsidy to a more expensive test.[9]

---

[8] Wiles suggests that his screening hypothesis (or what he calls external-test-not-content hypothesis) be checked by comparing the incomes of graduates in a given field of work having degrees in subjects more or less relevant to that field. The lower the correlation between income and relevance of degree subject, the more likely the hypothesis. Such an examination of incomes at a microlevel is certainly of high priority.

Conclusion

The screening hypothesis challenges the view that earnings differentials, standardized for differences due to other factors, measure the social returns to education. As formulated by Arrow (1973) the hypothesis is very difficult to test. If there are no social returns to the screening function of education, it could in principle be tested by production functions, but these are in practice beset with simultaneity problems. Nor do earnings functions for individuals provide a clear test, since the question is not *whether* education explains earnings, but *why* it does.

Most people would agree that screening is a part of the explanation, but how much of a part? The Taubman and Wales (1973) measure is not satisfactory. But three pieces of evidence suggest that screening is not a major part of the explanation. First, rates of return to dropouts are as high as to those who complete a course, which refutes the sheepskin version of the screening hypothesis. Second, standardized educational differentials rise with age, while a reasonable version of the screening hypothesis would lead one to expect a fall. Third, if screening is the main function of education, it could probably be done more cheaply by testing and other means, and agencies would have developed to reap the very large profits that could be made by doing this. Pending further work, we conclude that the theory of human capital is not after all in ruins.

Appendix 1

The Proportion of Educational Earnings Differentials Due to Education

The Appendix Table shows what proportion of crude intereducational earnings differentials survive after extracting the effects on earnings of other variables correlated with education but not caused by it. We do not allow here for any effect of screening, which constitutes round 2 of the debate referred to above; instead, we summarize the results of round 1. In each case, we compare the coefficient (b) obtained when an earnings variable is regressed on education and a minimum number of other "original" variables with the coefficient (b') obtained when earnings are regressed on education, the original variables, and a number of "additional" variables not caused by education. The table gives the values of b'/b. Where the earnings function includes interaction effects, the values of b'/b are shown for some representative group of people.

---

[9] Needless to say, even if screening were done by tests rather than education, the competitive equilibrium could well involve overinvestment. If types of labor were perfect substitutes (and ignoring the problem of incentives), it unquestionably would be.

APPENDIX TABLE

PROPORTION (b'/b) OF EARNINGS DIFFERENTIALS DUE TO EDUCATION (MALES, UNITED STATES)

		MAIN VARIABLES CONTROLLED FOR			
COURSE	b'/b	Original	Additional	SOURCE	SAMPLE
Higher education	65	Age	Ability, family background, marital status, health	Taubman and Wales (1973)	NBER sample of World War II veterans
B.A.	.97	Age	Ability	Hause (1972)	NBER sample
B.A.	.87	Age	Ability	Hause (1972)	Rogers (1969) (see below)
Years of schooling after military service	.88	Age, schooling before service, color, length of military service	Ability, father's status, place of schooling	Griliches and Mason (1972)	Bureau of the Census sample of postwar veterans
Years of schooling	.56	Age	Family size, father's occupation and income	Bowles (1972)	Mainly U.S. Census survey data
Years of schooling	.57	Age	IQ, father's occupation, and portion of ability and occupation not due to education	Jencks (1972)	Mainly March 1962 Current Population Survey data
B.A.	.86	Age	IQ, parental class, religion, marital status, no. of jobs	Rogers (1969)	Eighth graders in certain Connecticut schools, 1935
High school	.73	Age			
B.A.	.82	Age, race, region	Place of residence, hours worked, marital status, family size, region of birth	Hanoch (1965)	Census 1959

The striking finding is that the ratio is rarely less than Denison's (1964) assumed value of 3/5 and often higher. One must, however, recognize that many relevant variables such as motivation are not allowed for. If this is a major determinant of both earnings and educational choice, the figures in the table must greatly overstate the true position (*pace* Griliches 1970). We would not therefore consider that any figure approaching 100 percent of the crude age-standardized differentials is likely to be relevant to educational policy or to the analysis of the effects of education on growth.

Appendix 2

Interpreting the Coefficient on Schooling in an Earnings Function

This Appendix illustrates how the researcher's earnings function could show a positive effect of education ($g_S > 0$), even if he included in it the set of non-educational variables (A) which exactly determine productivity and if f_S is zero. For convenience we assume that A and A^E are each single variables (rather than vectors of attributes), positively but imperfectly correlated, and that all functions are linear. We then have the following representation of the general model outlined in the text (the general expression being shown in brackets):
Productivity relation:

$$P = a_0 + a_1 A \qquad [P = f(S, A)], \tag{1}$$

researcher's estimating relation:

$$Y = b_0 + b_1 S + b_2 A^R + u_2 \qquad [Y = g(S, A^R) + u], \tag{2}$$

employer's estimating relation:

$$P = c_0 + c_1 S + c_2 A^E + u_3 \qquad [P = h(S, A^E) + v], \tag{3}$$

employer's method of pay:

$$Y = \hat{P} = \hat{c}_0 + \hat{c}_1 S + \hat{c}_2 A^E. \tag{4}$$

Although schooling does not affect productivity ($f_s = 0$), it will affect income positively in the researcher's estimating equation ($\widehat{b_1} > 0$) if ($\hat{c}_1 S + \hat{c}_2 A^E$) is positively correlated with S, A held constant. If $\hat{c}_1$ and $\hat{c}_2$ are positive, sufficient conditions are that A^E be positively correlated with S, A held constant; in other words,

$$r_{SA^E} > r_{A^E A}\, r_{AS}.$$

Needless to say, the necessary conditions are more elaborate and less strong than this.

References

Arrow, K. "Higher Education as a Filter." *J. Public Econ.* 2, no. 3 (July 1973): 193-216.

Ashenfelter, O., and Mooney, J. D. "Graduate Education, Ability, and Earnings." *Rev. Econ. and Statis.* 50 (February 1968): 78-86.

Becker, G. S. *Human Capital.* New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1964.

Berg, I. *Education and Jobs: The Great Training Robbery.* Boston: Beacon, 1971.

Blaug, M. "The Correlation between Education and Earnings: What Does It Signify?" *Higher Education* 1, no. 1 (1972): 53–76.

Bowles, S. "Schooling and Inequality from Generation to Generation." *J.P.E.* 80, no. 3, suppl. (May/June 1972): S219–51.

Chiswick, B. R. "Schooling and Earnings of Low Achievers: Comment." *A.E.R.* 62 (September 1972): 752–54.

Denison, E. F. "Proportion of Income Differentials among Education Groups 'Due to' Additional Education: The Evidence from the Wolfe-Smith Survey." In *The Residual Factor and Economic Growth*. Paris: OECD, 1964.

Eckaus, R. "Returns to Education with Standardised Incomes." *Q.J.E.* 87 (February 1973): 121–31.

Fallon, P., and Layard, R. "Capital-Skill Complementarity, Income Distribution, and Output Accounting." *J.P.E.* (in press).

Griliches, Z. "Notes on the Role of Education in Production Functions and Growth Accounting." In *Education, Income, and Human Capital*, edited by W. Lee Hansen. New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1970.

Griliches, Z., and Mason, W. "Education, Income, and Ability." *J.P.E.* 80, no. 3, suppl. (May/June 1972): S74–103.

Hanoch, G. "Personal Earnings and Investment in Schooling." Ph.D. dissertation, Univ. Chicago, 1965.

———. "An Economic Analysis of Earnings and Schooling." *J. Hum. Resources* 2 (Summer 1967): 310–29.

Hansen, W. Lee. "Rates of Return to Investment in Schooling in the United States." *J.P.E.* 81, no. 2 (April 1963): 128–41.

Hansen, W. Lee; Weisbrod, B. A.; and Scanlon, W. J. "Schooling and Earnings of Low Achievers." *A.E.R.* 60 (June 1970): 409–18.

Hause, J. C. "Earnings Profile: Ability and Schooling." *J.P.E.* 80, no. 3, suppl. (May/June 1972): S108–38.

Jencks, C. *Inequality*. New York: Basic, 1972.

Johnson, G. E. "The Demand for Labor by Educational Category." *Southern Econ. J.* 37, no. 2 (October 1970): 190–204.

Layard, R.; Sargan, D.; Ager, M.; and Jones, D. *Qualified Manpower and Economic Performance*. New York: Penguin, 1971.

Masters, S., and Ribich, T. "Schooling and Earnings of Low Achievers." *A.E.R.* 62 (September 1972): 755–59.

Mincer, J. "Schooling, Experience, and Earnings." Mimeographed. New York: Nat. Bur. Econ. Res., 1972.

OECD. *Statistics of the Occupational and Educational Structure of the Labour Force in 53 Countries*. Paris: OECD, 1969.

Rogers, D. C. "Private Rates of Return to Education in the United States: A Case Study." *Yale Econ. Essays* 9 (Spring 1969): 89–134.

Spence, D. "Job Market Signaling." *Q.J.E.* 87, no. 3 (August 1973): 355–74.

Stiglitz, J. "The Theory of Screening, Education, and the Distribution of Income." *A.E.R.* (forthcoming).

Taubman, P. J., and Wales, T. J. "Higher Education, Mental Ability, and Screening." *J.P.E.* 81 (January/ February 1973): 28–55.

Tolley, G. S., and Olson, E. "The Interdependence between Income and Education." *J.P.E.* 79 (May/June 1971): 460–80.

Wiles, P. "The Correlation between Education and Earnings: The External-Test-Not-Content Hypothesis (ETNC)." *Higher Education*, vol. 3, no. 1 (1974).

Research and Development Activity and Profitability: A Distributed Lag Analysis

Ben Branch

Dartmouth College

There are at least three ways in which profits and R & D may be related. First, profits may influence subsequent R & D. Second, R & D may influence subsequent profits. And third, it is possible that R & D and profits are influenced simultaneously by some third factor. For example, government support or exogenous surges in demand could increase both at the same time. Although the first two of these relations involve influences of one variable on another over time, the problem of distinguishing between a simultaneous and a recursive relation has not been tackled in previous work. The current study employs distributed lag techniques with pooled time-series and cross-section data to discriminate among the alternative types of time relations. A sample of seven industries containing 111 firms covering the 1950–65 period was used to test these relations. The results indicate a tendency for R & D to influence future profitability and to be influenced by past profitability.

Introduction

A relationship of some sort between research and development expenditures (R & D) and profits has been posited, either explicitly or implicitly, in virtually every study of industrial R & D. Thus, Schumpeter (1950) maintained that monopoly power provides the firm with profits and security, which give it the opportunity to take the risks involved in R & D. Scherer (1965*a*) and Minasian (1962) found evidence supporting the hypothesis that R & D exerts an influence on subsequent profits, while Grabowski (1968) observed an association between profit rates and

---

I wish to thank F. M. Scherer for many valuable suggestions. Thanks are also due to Lee Baldwin, Wallace Gardner, Michael Klass, Martin Segal, and Lester Taylor, who have commented on various drafts. Responsibility for any remaining errors rests with the author.

[*Journal of Political Economy*, 1974, vol. 82, no. 5]

© 1974 by The University of Chicago. All rights reserved.

subsequent R & D. On the basis of machine tool industry evidence that managers increased R & D during slumps to stimulate demand, Brown (1957) concluded that R & D and profits may be inversely related over the business cycle (Brown 1957).

As the above examples suggest, there are at least three possible relations between profit and R & D. First, profits may influence subsequent R & D. Second, R & D may influence subsequent profits. And third, it is possible that R & D and profits are influenced simultaneously by some third factor. For example, government support or exogenous surges in demand could increase both at the same time.

Although the first two of these relations involve influences of one variable on another over time, the problem of distinguishing between a simultaneous and a recursive relation has not been tackled in previous works. Prior studies have been restricted to analyses of cross-section data for a single year or several years lumped together. In contrast, the current study employs distributed lag techniques with pooled time-series and cross-section data to discriminate among the alternative types of time relations.

The paper is divided into four sections. In Section I the hypotheses to be tested are explained. Section II discusses the data used. The models employed to test these hypotheses are considered in Section III. The results and conclusions are stated in Section IV.

I. The Hypotheses

When a firm considers undertaking or continuing a given R & D project, it probably has some idea of that project's costs and benefits. Estimates of the returns from R & D outlays are generally less certain than similar estimates for spending on plant and equipment or sales promotion. Companies which spend a larger-than-average proportion of their funds on R & D either have above-average opportunities or are less risk-averting than most firms.[1] By taking advantage of high-yield high-risk R & D opportunity, firms may increase their profits in two ways. First, profit relative to size may be increased by an above-average return on R & D projects. Second, R & D-created new products may permit increased growth. Thus total profits may increase even if profits relative to size do not.

The other main postulated relation is between profits and subsequent R & D. Research and development competes for funds with all other items of expenditure: production expenses, gross investment, advertising, interest, taxes, and dividends.[2] Many of these fund uses are unavoidable. Others, such as dividends, investment, and advertising, are likely, for strategic reasons, to be maintained at some above-minimal levels even when the sources of funds—after-tax profits, depreciation, and new securities issues—are constricted. Because of the well-known risks and relatively long time horizons associated with R & D, borrowing or the issuance of new equity securities is an unlikely source of funds for the support of R & D projects, and especially for projects at the margin of a firm's R & D portfolio. We might therefore expect changes in R & D expenditure normally to be positively associated with changes in profits.

---

[1] On the role of expected value maximization and risk aversion in R & D decisions, see Mansfield (1968).

However, in industries with low and cyclical profits and slow technological change, the assumption that residual funds to support R & D become available in greater abundance when profits are high may not hold. Low and cyclical profits are likely to shorten management's time horizon, while a history of slow technological change is associated with managers who consider R & D expendable.[3] In cyclical industries, high profits usually coincide with heavy demand and operations at a high level of capacity. In such cases profits may be inversely associated with R & D for three reasons. First, as managers attempt to adjust investment to booming demand, R & D, with a typically long wait until payoff, may be robbed of funds to finance capital expansion. Second, in periods of high-capacity operation the firm's staff of scientists and engineers may be shifted from R & D to routine product engineering and to solving the everyday problems brought on by overuse of facilities. Third, managers may be satisfied with the status quo when profits are high, while they may seek to stimulate demand through R & D when profits are low.[4]

Now that the two basic hypotheses have been considered separately, it is essential to explore relations between them. In a combined cross-section–time-series sample there are two types of variation: variation over time and over the cross-section. A correlation between firms' *average* R & D intensity and profitability will be referred to as the cross-section relation, while a correlation which depends on the *time lag* between R & D intensity and profitability will be referred to as the time-series relation. If profits and R & D intensity are determined simultaneously, a cross-section relation may be found. Simultaneous causation may also produce a time-series relation. If the third factor causing the simultaneous relation has a time pattern, and if the lags between this factor and profits and R & D intensity are equal, the correlation should be highest when there is no lag between them. Since our hypotheses suggest that above-average R & D intensity and above-average profits tend to reinforce one another, a two-way recursive system (implying that both hypotheses are correct) may also produce a cross-section relation. However, in such a system R & D intensity is expected to be more highly correlated with past and future profits than with the current value. It is also possible for a given industry that both simultaneous and recursive relations are present. The tests which follow will seek to determine which time pattern of relationships is most compatible with the available evidence.

---

[2] While R & D may be risky, it is also risky for the enterprise in an industry with rapid technological change to do so little R & D that its technology becomes obsolescent. For such a firm, the expected return from doing sufficient R & D to avoid becoming technologically obsolete is generally quite high. Thus, even if the firm behaves as a risk averter, it will typically endeavor to do sufficient R & D to avoid being put out of business by technologically advanced rivals.

[3] For an interesting study of the interrelated nature of firms' decisions with regard to investment, advertising, dividends, and R & D, see Mueller (1967).

[4] Consistent with this view are the findings of Scherer (1965*b*) and Baxter and Grabowski (1967) that the variance of R & D intensity was much higher in industries with low R & D than in industries with high R & D. Thus managers in industries with high R & D intensity imitate their rivals more closely than managers in industries having low R & D intensity.

II. The Data

Data from a sample consisting of observations on 111 firms for the 1950–65 period were used to explore the hypotheses developed thus far.

The companies were selected on the following basis. Those on the FTC's list of the 1,000 largest manufacturers in 1950 that survived as separate firms in 1969 were classified as to principal industry.[5] For each industry, the sample consisted of those firms that were not excluded for one of the reasons mentioned below and which were primarily occupied in one of the seven industries listed in table 1.[6]

Of the original 1,000 on the 1950 FTC list, 468 firms were eliminated because they had gone out of business, been acquired by another firm, participated in a merger which substantially changed their identity, or were private firms not publishing formal data. Of the remaining 532, 157 were classified in the seven industries of table 1. Forty-six of these firms had to be excluded because consistent data were unavailable, because they were subject to some degree of government regulation, or because they did a major fraction of their work under government contract. It would have been possible to extend the sample to additional industries. For most of the remaining industries, however, one would not expect to find much relation between profits and R & D for one or more of the following reasons: an insignificant amount of R & D is conducted (food and clothing); most R & D work is done under government contract (aircraft); or the industry has an institutional feature which blurs any relation (the patent pooling policy of the auto industry). Thus,

---

[5] See Brown for an analysis of this phenomenon in the machine tool industry (1957).

[6] The groupings were made on the basis of the SEC classification of firms filing annual reports. See U.S. Securities and Exchange Commission (1966).

TABLE 1

THE 1950–64 SAMPLE

Industry	SIC	No. of Firms in Sample	Description
Chemicals	281	13	Industrial organic and inorganic chemicals, plastic materials and synthetic resins, synthetic rubber and other manmade fibers except glass
Electrical equipment	357	22	Office computing and accounting equipment
	361	...	Electrical transmission and distribution equipment, electrical industrial apparatus, lighting and wiring equipment, miscellaneous electrical machinery and equipment
	363	...	Household appliances
	365	...	Radio and television receiving sets
Paper	262	18	Paper and allied products
Mechanical equipment	352	28	Farm machinery; construction, mining, and material-handling equipment
	351	...	Engines and turbines
	354	...	Metalworking machinery and equipment
	355	...	Special industrial machinery except metalworking
	356	...	General industrial machinery and equipment
	358	...	Service industry machinery
Nonferrous metals	333	11	Smelting and refining of nonferrous metals including rolling, extruding, and nonferrous foundries
Petroleum refining	291	10	Petroleum refining
Drugs	283	9	Drugs

while 111 firms is a rather small subsample of the 1,000 largest firms in 1950, it represents a much larger fraction of the firms for which the data are good enough to expect on *a priori* grounds a relation between R & D and profits.

Mergers consummated during the period of the study and large enough to be recorded by the FTC were treated as having occurred at the beginning of the period. That is, the data for such firms were consolidated from 1950 forward, rather than from the actual date of the merger. Smaller mergers were ignored.

Because no other variables can be measured in such rich abundance and because of its high correlation with other indexes of inventive activity (R & D spending, R & D employment, count of significant inventions, or sales of new products),[7] a count of the number of patents each firm received in a given year was used as the measure of its R & D activity. For the 1950–65 period there was a lag of approximately 4 years between reducing an innovation to practice and receiving a patent on it.[8] For this reason the patents received in year t are used as an index of a firm's R & D output in year $t - 4$.

---

[7] It has been shown that the number of patents a firm receives is highly correlated with other indexes of inventive activity. See Mueller (1966) and Comanor and Scherer (1969).

Patent data have several well-known limitations. First, the propensity to patent has not been constant over time (see Schmookler 1966, pp. 28–30). Second, the lag between reducing an innovation to practice and patent issue is not the same for each innovation but varies over time and over the cross section of patents. Third, the propensity to patent varies from industry to industry and from firm to firm. Fourth, the quality of innovations on which patents are issued varies enormously.

All of the above-mentioned shortcomings reflect errors in measurement. Previous work (Mueller 1966; Comanor and Scherer 1969) has demonstrated that these errors are not large enough to overwhelm the basic phenomena for which they are an index. The problems imposed by random errors in measurement may be reduced by increasing the size of the sample. Difficulties caused by differences in the propensity to patent and time lag may be reduced by working with individual industries which, presumably, have similar time lags and propensities to patent. There are of course industries in which little R & D is done and thus few patents are received, but these have been excluded from the sample.

Even in industries that contain only firms which receive large numbers of patents each year, the problems of errors in measurement do not disappear altogether. When there are random errors in measurement in the independent variable, as there are for the data used to test the "R & D leads to profits" hypothesis, the coefficient of the R & D intensity variable is biased downward. This both causes an understatement of the true R & D coefficient and increases the likelihood of rejecting a correct hypothesis. Random errors in the dependent variable, as there are in the "profits lead to R & D" hypothesis, increase the standard errors of the independent variables, again increasing the likelihood that a true hypothesis will be rejected. Thus, if we find a relation where we have random errors of measurement in an independent or dependent variable, it is likely that we have done so in spite of, not because of, these errors.

The period 1950–65 was selected because the most recent patent data available were for 1969 and, allowing for the 4-year lag, 1965 was the latest year for which our index of R & D activity was available;[9] 1950 was selected as the beginning date, since the post–World War II readjustment had largely ended by that time. Thus, the sample consists of companies in seven industries[10] for which data were available over a 17-year period. These data are pooled, giving a combined cross-section–time-series sample with an observation for each firm by year.

---

[8] This lag between reducing to practice and patent issue is the sum of two lags: an invention-to-application lag and an application-to-issue lag. Sanders (1962) gives 9 months as the average invention-to-application lag. Using a random sample of 111 patents, Scherer (1965*b*) found a mean application-to-issue lag of 3.5 years for the 1950s, with a standard deviation of 1.6 years.

[9] That is, patent data were collected for the 1950–65 period. Financial data for the firms were collected for the 1947–69 period. In this way it was possible to lag the R & D variables without losing more than 1 year of observation.

III. The Models

As a result of differences in resources (human and nonhuman) and unexpected chances in demand, firms are likely to differ in their profitability.[11] Over time, a firm's profits are likely to be influenced by the business cycle and past R & D. This suggests the following model:[12]

$$\begin{aligned} r_{it} = {} & \Phi_i FD_i + \rho I_t \\ & + \omega\left(\varepsilon \frac{\hat{P}_{it}}{A_{it}} + \frac{\hat{P}_{it-1}}{A_{it-1}} + \frac{\eta \hat{P}_{it-2}}{A_{it-2}} + \frac{\eta^2 \hat{P}_{it-3}}{A_{it-3}} \cdots + \frac{\eta^{n-1} \hat{P}_{it-n}}{A_{it-n}} \cdots\right), \end{aligned} \tag{1}$$

where r = profits after taxes plus interest paid dividend by assets,[13] P = the number of patents received by the firm; A = the assets of the firm, in millions of dollars; FD = firm dummy having a value of one for the ith firm and zero otherwise; I_t = the percentage deviation from trend of the industry's Federal Reserve Board production index;[14] $\hat{P}$ = P

---

[10] Data for an eighth industry (ferrous metals) were also collected. The drastic change in profitability for this industry from the 1950s to the 1960s led to its deletion.

[11] In preliminary work, an attempt was made to explain these interfirm differences by putting in variables (size, capital structure, etc.) other than firm dummies. They were not sufficient to remove the cross-section relation. This caused the lagged dependent variable to be biased upward.

[12] Profits and R & D are collinear because of their relation to a firm's size. To avoid the unstable coefficients which could result from fitting a levels equation, the models are specified in a ratio form. For a discussion of the appropriateness of this approach, see Kuh and Meyer (1955). The ratio of patents received to the firm's assets is used as an index of R & D intensity. Other size deflators could have been used. It is unlikely that they would have changed the results significantly. A few regressions were run using net worth as the deflator. It made very little difference. Assets were chosen over other possible deflators such as sales because assets tend to fluctuate less over the business cycle. This specification implies a linear relation between size and R & D activity. This relation was tested for each of the seven industries by regressing each firms average $(\overline{P/A})$ on a constant and $\bar{A}$, $\bar{A}^2$, $\bar{A} + \bar{A}^2$, Ln $\bar{A}$, Exp $\bar{A}$. For only one of the industries (petroleum) are any of these variables significant at the 90 percent level. The results for petroleum were not significant at the 95 percent level. Thus it was felt that the ratio form did not unduly bias the results.

[13] Profit after taxes plus interest is an index of the firms return on its resources. To be consistent, both R & D activity and profits plus interest are deflated by assets.

[14] Had earnings been expressed in a levels form, then there would have been no need to detrend the index. With earnings in a ratio form, however, detrending becomes necessary.

corrected for differing Patent Office resources over time;[15] and $\tau = t - y$ where y is a nonnegative integer determined by fitting (1).

The lag structure for the R & D intensity variables rises to a peak and declines geometrically thereafter. An example of such a lag structure is presented in figure 1. While there is no strong reason for expecting a geometric increase (as opposed to a linear increase), this assumption was made to permit the use of a Koyck transformation. With a rapid decay rate (low value of η) there is little difference between the geometric decline assumed and any other likely decline form.

Applying the Koyck transformation to equation (1) yields

$$r_{it} = \Phi_i(1 - \eta)FD_t + \rho(I_t - \eta I_{t-1}) + \omega\left[\varepsilon \frac{P_{i\tau}}{A_{i\tau}} + (1 - \varepsilon\eta)\frac{P_{i\tau-1}}{A_{i\tau-1}}\right] + \eta r_{it-1}. \qquad (2)$$

This is the equation fitted for each of the seven industries. An approximate value for η was found through iteration. That is, an estimated value for η was used to specify $(I - \eta I_{t-1})$, and the regression was run. The resulting coefficient η for $P_{i\tau-1}/A_{i\tau-1}$ was then substituted for the original estimate and the regression run again. This process was repeated until the η's were approximately equal. Various values for ε and τ were tried, and the one which produced the highest t value for ω was selected as the best estimate.

For the hypothesis that R & D leads to sales growth, the following model was used:

$$G_{it} = \alpha FD_t + \beta I_t + \gamma\left[(1 - a)\frac{P_{i\tau}}{S_{i\tau}} + a\frac{P_{i\tau-1}}{S_{i\tau-1}}\right], \qquad (3)$$

where $G_{it} = (\text{Sales}_{it} - \text{Sales}_{it-1})/\text{Sales}_{it}$.

Since the dependent variable involves current sales, consistency requires that patents also be deflated by sales. A long-period distributed lag such as is implied by the Koyck transformation was not used, as this would imply that R & D leads to continually increasing sales. In fact, if a given amount of R & D raises sales in t, this may increase the difficulty of further sales growth in $t + 1$ as the base is higher.

We must also explore the opposite direction of causality. We expect company R & D, measured by appropriately lagged patents, to vary

---

[15] The Patent Office correction was obtained as follows. The total number of patents issued from 1954 to 1969 was regressed on e and pk, where e is Patent Office employment and pk measures the printing budget of the Patent Office in constant dollars. Next the trend rates for e and pk (determined by regressing the natural logarithm of Patent Office employment and its printing budget in constant terms) were found ($\hat{e}$ and $\hat{p}k$). Then the difference in the predicted value of total patents was determined if $\hat{e}$ and $\hat{p}k$ are used rather than e and pk. The correction equals actual patents minus this difference divided by the actual value.

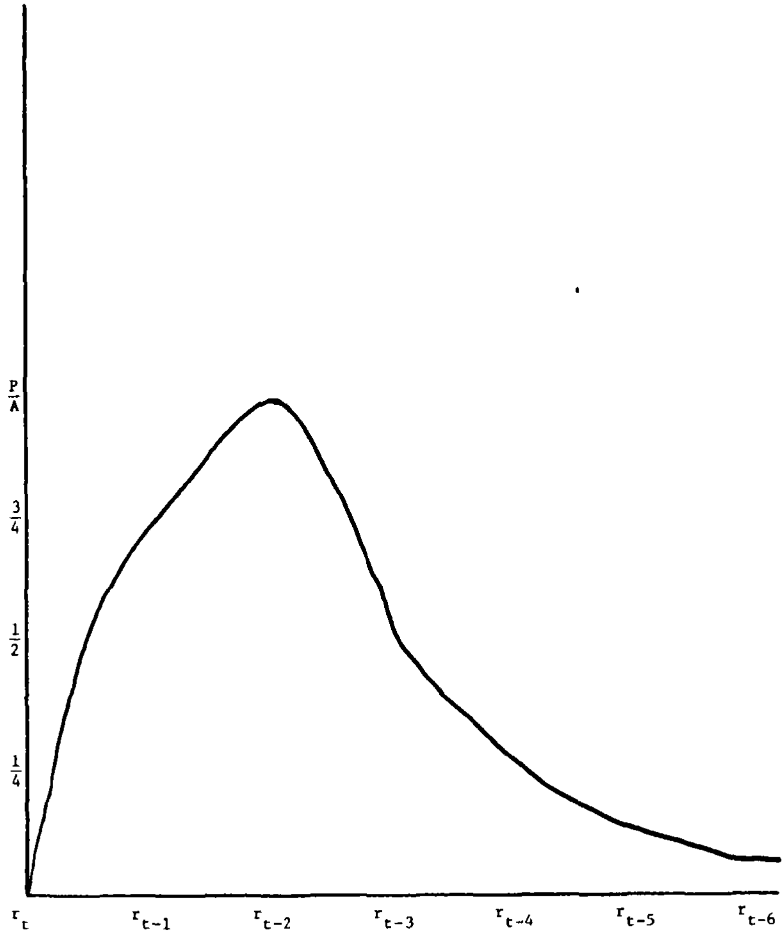


FIG. 1.—Lag structure for eq. (5) with peak at $t - 2$. $\varepsilon = .75$; $\eta = .5$.

with interfirm differences in opportunity and risk aversion and with the availability of funds for nonessential projects. Since the number of patents issued by the Patent Office in a given year is influenced by the Office's resources, some correction must be made for the effect of resource fluctuations. In years of tight budgetary constraint, the Patent Office sometimes hoards its skilled patent examiners while cutting its printing budget. In other years its labor force is reduced. For this reason both factors are introduced into a regression explaining company patenting. Since R & D projects often take several years to complete, it is expected that the number of patents received in a given year is influenced by profit in several previous years, with declining weights after some initial peak. These considerations suggest the following model:

$$\begin{aligned}\frac{P_{it}}{A_{it}} = \alpha_i(FD_i) + \beta E_t + \delta PK_t \\ \gamma(\theta\pi_{t-1} + \pi_{it-2} + \lambda\pi_{it-3} + \lambda^2\pi_{it-4} + \dots),\end{aligned} \tag{4}$$

where E = the percentage deviation from trend of Patent Office employment; PK = the percentage deviation from trend of the Patent Office's printing budget (in constant dollars); and π = the firm's profits after taxes, divided by assets. Performing a Koyck transformation on equation (4) yields

$$\begin{aligned}\frac{P_{it}}{A_{it}} = \alpha_i(1 - \gamma)FD_i + \beta(E_t - \lambda E_{t-1}) + \delta(PK_t - \lambda PK_{t-1}) \\ + \gamma[\theta\pi_{it-1} + (1 - \gamma\theta)\pi_{it-2}] + \lambda \frac{P_{it-1}}{A_{it-1}}.\end{aligned} \tag{5}$$

This equation was fitted by the use of iterative techniques similar to those used with equation (2).

Equations (2) and (5) together constitute a dynamic system. This system will be examined to determine conditions under which it is stable. A look at the data reveals that interest paid is very small relative to profit after taxes. Thus r and π are assumed to be equal. We would like to know what the ultimate impact on r is for an initial change Δr. The direct effect (that is, the effect without considering any interactions between r and P/A) of a change Δr is to change P/A by $\gamma\Delta r$. The secondary effect of this change in P/A is to change r by $\omega(\gamma\Delta r)$. Thus, an initial change of r leads to a further change of $\omega\gamma\Delta r$; this in time leads to a change of $(\omega\gamma)^2$, and so on. The ultimate impact is $\Delta r + \omega\gamma\Delta r + (\omega\gamma)^2\Delta r + (\omega\gamma)^3\Delta rt \dots$ The sum of this infinite series is $\Delta r/(1 - \omega\gamma)$. This sum is finite for $\omega\gamma < 1$.

In like manner, an initial change in P/A of $\Delta P/A$ leads to an ultimate change of $(\Delta P/A)/(1 - \omega\gamma)$. Thus, the system will be stable when $\omega\gamma < 1$.

IV. Results and Conclusions

A summary of the results is presented in tables 2, 3, and 4. These results strongly support the hypothesis that R & D activity tends to increase both profits and growth. The two hypotheses are accepted at the 90% level or better for each of the seven industries.[16] For six of the seven industries

---

[16] In all cases the coefficient of the best lag structure significantly outperformed any other lag structure beginning at a different period. A check of the residuals did not reveal any systematic bias. Thus the use of firm dummies and a ratio form appears to have eliminated any serious problem with serial correlation and heteroscedasticity.

TABLE 2

R & D LEADS TO PROFITS

	COEFFICIENT FOR:					
INDUSTRY	R & D (ω)	Lagged Profit (R)	Index (ρ)	BEGINNING OF LAG (y)	WEIGHT OF FIRST YEAR (ε)	$\bar{R}^2$
Chemical	0.021*** (4.0)	0.316*** (5.03)	0.1026*** (3.30)	1	1.34	.830
Electrical equipment ...	0.008** (2.03)	0.572*** (13.30)	0.101*** (2.76)	2	1.75	.740
Paper	0.017*** (3.71)	0.086 (1.50)	0.288*** (3.20)	1	11.62	.522
Mechanical equipment ...	0.017* (1.94)	0.340*** (7.50)	0.100*** (5.74)	3	0.99	.504
Nonferrous metals	0.083* (1.73)	0.421*** (7.47)	0.152*** (9.78)	1	0.16	.721
Petroleum	0.056** (2.59)	0.635*** (11.9)	0.035*** (7.68)	1	0.47	.881
Drugs	0.051*** (3.23)	0.686*** (11.08)	0.013 (0.24)	0	0.13	.875

NOTE.—t-values are given in parentheses
* Significant at 90% level.
** Significant at 95% level.
*** Significant at 99% level.

TABLE 3

R & D LEADS TO SALES GROWTH

	COEFFICIENTS FOR:				
INDUSTRY	R & D (γ)	Index (β)	BEGINNING OF LAG (y)	WEIGHT OF FOLLOWING YEAR	$\bar{R}^2$
Chemicals	0.270*** (3.24)	0.933*** (4.59)	1	0.2	.115
Electrical equipment	0.136** (2.07)	0.209*** (2.85)	1	0	.066
Paper	0.497*** (3.13)	1.222*** (7.09)	1	0.3	.191
Mechanical equipment	0.182* (1.93)	0.559*** (7.42)	2	0.5	.095
Nonferrous metal	1.185*** (3.43)	0.488*** (5.58)	1	0	.162
Petroleum	0.316** (2.27)	0.955*** (6.16)	1	0.8	.235
Drugs	0.144* (1.89)	0.292 (1.23)	3	0	.082

NOTE.—t-values are given in parentheses
* Significant at 90% level.
** Significant at 95% level.
*** Significant at 99% level.

TABLE 4

PROFIT LEADS TO R & D

INDUSTRY	COEFFICIENTS FOR: Lagged R & D (γ)	Profit (λ)	Employment of Patent Office (β)	Printing Budget of Patent Office (δ)	BEGINNING OF LAG (y)	WEIGHT OF FIRST YEAR (θ)	$\bar{R}^2$
Chemical . . .	−0.167 (0.62)	0.664*** (11.01)	1.134*** (4.42)	0.292*** (2.82)	2	1.50	.634
Electrical equipment	0.411*** (2.85)	0.734*** (18.06)	0.322 (1.49)	0.368*** (4.16)	2	1.36	.680
Paper	0.134** (2.2)	0.171*** (2.58)	0.029 (0.26)	0.009 (0.17)	1	0.6	.590
Mechanical equipment	0.233** (2.33)	0.617*** (16.6)	0.274* (1.84)	0.051 (0.83)	4	1.62	.614
Nonferrous metal . .	0.014 (0.44)	0.364*** (4.01)	−0.026 (−0.28)	0.044 (1.16)	2	2.75	.777
Petroleum .	0.330*** (3.23)	0.462*** (4.29)	0.020 (0.44)	0.092*** (2.96)	1	2.16	.801
Drugs	0.036 (0.37)	0.322*** (3.78)	0.244* (1.87)	0.033 (0.009)	1	3.10	.678

NOTE.—t-values are given in parentheses.
* Significant at 90% level.
** Significant at 95% level.
*** Significant at 99% level.

the peak impact of R & D activity on sales growth either precedes or is simultaneous with the peak impact on profits. Only the drug industry is an exception. Perhaps this is a result of drug companies introducing new products with very high margins, with competition forcing a reduction in prices, which in turn causes an increase in sales and a decrease in profits.

While not as uniform as the results of tables 2 and 3, the "profit leads to R & D" hypothesis is also given some support by the results. The hypothesis is accepted at the 95% level for four of the industries. For chemicals the sign is negative and insignificant. Nonferrous metals has the expected sign, but the coefficient is insignificant. Possibly nonferrous metals corresponds to some extent to Brown's (1957) postulated behavior. That is, above-average profits reflect an increase in both the availability of funds and the need for such funds for non–R & D purposes. In the case of drugs it is possible that a high average profit rate reduces the constraining effect fluctuating profits might otherwise have.

The somewhat weaker support found for the "profit stimulates R & D" hypothesis may be due to its dependence on managerial behavior. That is, managers must react to higher profits by increasing R & D spending for the "profit leads to R & D" relation to hold, while the other direction of causation exists if R & D spending is of above-average profitability.

References

Baxter, Nevins D., and Grabowski, Henry G. "Imitation, Uncertainty, and the Research and Development Decision." Paper presented at Conference of the Institute of Management Science, April 6, 1967.

Brown, William H. "Innovation in the Machine Tool Industry." *Q.J.E.* 71, no. 3 (August 1957): 406–25.

Comanor, W. S., and Scherer, F. M. "Patent Statistics as a Measure of Technical Change." *J.P.E.* 77, no. 3 (May/June 1969): 392–98.

Grabowski, Henry G. "The Determinants of Industrial Research and Development: A Study of the Chemical, Drug, and Petroleum Industries." *J.P.E.* 76, no. 2 (March/April 1968): 292–306.

Kuh, E., and Meyer, J. "Correlation and Regression Estimates When the Data Are Ratios." *Econometrica* 23, no. 4 (October 1955): 400–416.

Mansfield, Edwin. *Industrial Research and Technological Innovation: An Econometric Analysis.* New York: Norton, 1968.

Minasian, Jora R. "The Economics of Research and Development." In *The Rate and Direction of Economic Activity.* Princeton, N.J.: Princeton Univ. Press (for Nat. Bur. Econ. Res.), 1962.

Mueller, Dennis. "Patents, Research and Development and the Measurement of Inventive Activity." *J. Indus. Econ.* 15, no. 1 (November 1966): 28–37.

———. "The Firm Decision Process: An Econometric Investigation," *Q. J* (February 1967): 58–87.

Sanders, Barkev S. "Some Difficulties in Measuring Inventive Activity." I *Rate and Direction of Economic Activity.* Princeton, N.J.: Princeton Univ (for Nat. Bur. Econ. Res.), 1962.

Scherer, Frederic M. "Corporate Inventive Output, Profit and Growth." J 73, no. 3 (June 1965): 290–97 (*a*)

———. "Firm Size, Market Structure, Opportunity, and the Output of Patented Inventions." *A.E.R.* 55, no. 1 (December 1965): 1097–1125. (*b*)

Schmookler, Jacob. *Invention and Economic Growth.* Cambridge, Mass.: Harvard Univ. Press, 1966.

Schumpeter, Joseph A. *Capitalism, Socialism, and Democracy.* New York: Harper & Row, 1950.

U.S., Securities and Exchange Commission. *Directory of Companies Filing Annual Reports with the Securities and Exchange Commission.* Washington: Government Printing Office, 1966.

Communications

A Note on Border-Tax Adjustments[1]

James E. Meade

Cambridge University

If a country imposes a 10 percent duty on all imports and a 10 percent susbsidy on all exports, this is equivalent to a 10 percent devaluation of its currency insofar as its commodity trade is concerned. This will have an effect on the balance of trade unless and until it is offset by a 10 percent revaluation of the currency or a general 10 percent inflation of all domestic money prices and costs. It will then have no effect upon quantities of production, consumption, import, or export of any commodity. Its only effect will be a monetary one on the rate of exchange between the domestic currency and foreign currencies or on the general level of prices and costs in terms of the domestic currency.

This well-known proposition can be extended to cover a group of countries. Consider three countries—A, B, and C. Suppose that A and B (having perhaps formed an economic union) agree to impose a 10 percent duty on imports from C and a 10 percent subsidy on exports to C without taxing or subsidizing their imports from and exports to each other. If A's and B's currency both appreciate by 10 percent in terms of C's currency, or if A's and B's domestic money price and cost levels both go up by 10 percent relatively to C's domestic money prices and costs, there will be no real effect on any quantity of any commodity produced, consumed, or traded.

The proposition can then be further extended by combining the two ideas expressed in the preceding two paragraphs. Imagine the following sequence of events.

1. First A imposes a border-tax adjustment of t_{bc} on its trade with B

---

[1] This note is an attempt to give a general formula for a proposition about border-tax adjustments to be found at the end of section 1 of Johnson and Krauss (1970).

[*Journal of Political Economy*, 1974, vol. 82, no. 5]

© 1974 by The University of Chicago. All rights reserved.

and C, where t_{bc} is the ad valorem value of a tax imposed by A on its imports from B and C and also the ad valorem value of a subsidy paid on A's export to B and C.

2. B then imposes a border-tax adjustment of t_{ac} on its trade with A and C.
3. C then imposes a border-tax adjustment of t_{ab} on its trade with A and B.
4. A and B then get together and impose a border-tax adjustment of t_c on their trade with C in addition to the border-tax adjustments of t_{bc} and t_{ac} which each has already independently imposed on its imports from all sources and its exports to all sources.
5. A and C then get together and impose an additional border-tax adjustment of t_b on their trade with B.
6. Finally, B and C get together and impose an additional border-tax adjustment of t_a on their trade with A.

In accordance with the analysis given in the first two paragraphs of this note, none of these events introduces any real distortion of trade, though each may require an adjustment of exchange rates or of domestic price levels to maintain balances of trade. The combined effect of events 1–6 will then be to leave trade undistorted.

Now let ${}_aT_b$ represent the actual combined border-tax adjustment which A will finally be imposing on its trade with B. We have

$$ {}_aT_b = t_{bc} + t_b, $$

and similarly

$$ \begin{aligned} {}_aT_c &= t_{bc} + t_c, \\ {}_bT_a &= t_{ac} + t_a, \\ {}_bT_c &= t_{ac} + t_c, \\ {}_cT_a &= t_{ab} + t_a, \\ {}_cT_b &= t_{ab} + t_b. \end{aligned} \tag{1} $$

From the six equations in (1), one can eliminate the six particular border-tax adjustments—the six small t's—and obtain the formula

$$ {}_aT_b + {}_bT_c + {}_cT_a = {}_bT_a + {}_cT_b + {}_aT_c. \tag{2} $$

This is the general formula for a set of border-tax adjustments that does not distort trade.

The argument has so far been expressed in terms of a three-country world. But the formula in equation (2) can in fact be applied to an N-country world without any modification. Consider, for example, a five-country world comprising A, B, C, D, and E. Let t_{bcde} be border-tax adjustment that A makes on its trade with B, C, D, and E. Let t_{bcd} be the border-tax adjustment that A and E jointly agree to make on their trade with B, C, and D. Let t_{bc} be the border-tax adjustment that A, D, and E jointly agree to impose on their trade with B and C. Let t_b be the border-tax adjustment that A, C, D, and E jointly agree to impose on their trade with B, and similarly for all the other t's.

Then the actual border-tax adjustment that A will impose on its trade with B can be expressed as:

$${}_aT_b = t_{bcde} + t_{bcd} + t_{bde} + t_{bce} + t_{bc} + t_{bd} + t_{be} + t_b.$$

If one writes similar expressions for ${}_bT_c$, ${}_cT_a$, ${}_bT_a$, ${}_cT_b$, and ${}_aT_c$, one finds once again, as in equation (2), that

$${}_aT_b + {}_bT_c + {}_cT_a = {}_bT_a + {}_cT_b + {}_aT_c.$$

The present author is not a sophisticated mathematician, but the above demonstration is sufficient to satisfy him that the formula in equation (2) is a perfectly general one. In other words, if all trade distortion is to be avoided, the actual border-tax adjustments made by any set of three countries to the trade among them must satisfy the formula in equation (2). Any world system that satisfies this general rule will not distort trade, though it will, of course, affect the exchange rates and/or domestic price levels needed to bring particular balances of trade into equilibrium.

The meaning of the general formula in equation (2) can be readily understood by analogy with the theory of cross-rates in foreign excha markets. When A imposes a border-tax adjustment of ${}_aT_b$ on its tr with B, this is equivalent to a devaluation by ${}_aT_b$ of A's currency in te of B's currency; and when B then imposes a border-tax adjustmen ${}_bT_a$ on its trade with A, this is equivalent to a devaluation by ${}_bT_a$ of currency in terms of A's currency. The net effect is the equivalent depreciation of A's currency in terms of B's currency by $({}_aT_b - {}_bT_a)$ Similarly, $({}_aT_c - {}_cT_a)$ is the equivalent of a net depreciation of A's currency in terms of C's currency, and $({}_cT_b - {}_bT_c)$ is the equivalent of a net depreciation of C's currency in terms of B's currency. But in order that cross-rates be in equilibrium, the net depreciation of A's currency in terms of B's currency must be equal to the sum of the net depreciation of A's currency in terms of C's currency and the net depreciation of C's currency in terms of B's currency; for equilibrium we must have

$${}_aT_b - {}_bT_a = ({}_aT_c - {}_cT_a) + ({}_cT_b - {}_bT_c),$$

which gives the same result as equation (2).

Reference

Johnson, Harry G., and Krauss, Melvyn B. "Border Taxes, Border Tax Adjustments, Comparative Advantage, and the Balance of Payments." *Can. J. Econ.* (November 1970). Reprinted in *The Economics of Integration*, edited by Melvyn B. Krauss. London: Allen & Unwin, 1973.

Policy Influence without Policy Choice

Donald C. Cell

Cornell College

This note attempts to explain how unorganized, passive, ill-informed voters collectively influence policy choice without expressing policy preferences. In our theorizing about democracy we have thought too exclusively of the voter as choosing between candidates and parties according to his policy preferences, like a consumer choosing between alternative sets of products. Consequently we have too often assumed that the electorate must somehow discriminate between *proposed policy alternatives* in order to influence policy choice.[1]

Toward a broader theory of voter influence, we shall treat *actual policy decisions* as governmental inputs into the voter's environment that, together with his own choices, affect his ex post utility.[2] In this model, voters greatly economize their political efforts by voting to remove government whenever they become dissatisfied with the utility cha they attribute to it; voters judge performance without understanding

---

Among those who gave valuable help on drafts of this paper, I especially thank H Park, William Hendley, Carl Shoup, Robert Dorfman, and the referees.

[1] What I call the policy preference theory of candidates and parties construes elec as primarily the first stage of the policymaking process through which voters broadly determine how the policy goals and instruments that are of interest to them will be ranked in the second or legislative stage. The rigorous statement of this participatory voting theory by Anthony Downs (1957) stimulated the present paper. Downs argued that voters cannot influence policy unless they (1) consider alternative policies and then (2) communicate and (3) constrain government to implement their specific ex ante policy preferences, because if they delegate these decisions they will lose control over them (pp. 105–9, 233–34, 247–52). Downs recognized and even briefly supported the counter-possibility of voter influence through ex post performance rating but then dropped it (pp. 41–45). He concluded that, aside from the understandable influence of organized pressure groups, the very rationality that keeps individual voters uninformed about policy alternatives conflicts with the social goal of electoral influence (pp. 127, 135–39, 244–47). One frequently encounters such doubts about the scope of voter influence that seem to be based upon implied versions of the policy preference theory of voting. See, for example, Becker (1958), Olson (1965, p. 165–67), Tulloch (1967, p. 120–22), Steiner (1970, pp. 46–49), Due and Friedlaender (1973, pp. 59–62). The policy preference theory is, indeed, consistent with the extreme rule-by-pressure-groups thesis of the pluralist tradition.

[2] Becker's version (1971, pp. 42–48) of the new consumer theory indicates how government policies that do not enter the voter's preference system affect his choice opportunities by changing his personal "production functions" for more fundamental commodities.

[*Journal of Political Economy*, 1974, vol. 82, no. 5]
© 1974 by The University of Chicago. All rights reserved.

foregone alternatives by applying their often crude ex post utility standards. The more voters so economize, the more they delegate the choice of policy goals without relinquishing influence over policy results. As compared to the voter's meager incentive, it turns out that the professional politician's stock in trade is his expertise in forecasting which policies will please voters.

Section I explains why voters put little effort into forecasting the utility consequences of candidates and parties. Section II develops an ex post evaluation model in which voters influence without forecasting. Section III explains why aspects of elections (candidate unreliability, inefficient communication, and candidate convergence) that make it difficult for voters to express ex ante preferences are not the barriers to voter influence they seem to be.

I. Policy Information as a Public Good

True, elections do offer an opportunity to participate collectively in policy choice in the process of selecting a government. But before we say anything about the individual's political intentions, we should first ask why he makes any political effort at all.

In his classic critique of participatory democratic theory, Schumpeter (1950, pp. 256–64) argued that the citizen loses his sense of reality and responsibility when he moves into mass politics because he leaves behind the favorable and unfavorable experience that disciplines his personal decisions. Public-good theory has now shown more precisely that the voter suffers no illusion when he fails to see a connection between his own contribution to mass politics and the policy decisions of government; for the single voter there simply *is no such connection*, since in all large constituencies neither the voter's utility from government nor the social performance of government will change if this one voter decides to withdraw his political efforts. But if the government's performance does indeed depend on some collective contribution by the electorate, it follows that this contribution is a public good. So the electoral process, it seems, requires the supply of a public good by individuals. Why, then, does the very individualism that electoral democracy expresses not also undermine it?

Since many people do first *decide how to vote* and then second *make the journey to the polls*, they must be getting satisfaction from these complementary efforts in two respects: (1) though the collective effect on government performance is a public good, acts such as deciding and going to the polls are private goods in themselves (like, for example, the nonmonetary satisfactions of a job); (2) people do have private demand schedules that internalize some of the benefits of public goods (charities, professional organizations like AAUP, environmental amenities, political parties, etc.), though we expect this demand usually to be highly elastic to cost.

Thus the economic explanation of large-scale voting emphasizes the small costs of the journey to the polls (Downs 1957, p. 269; Olson 1965, p. 164, n.; Tulloch 1967, pp. 108–14). By the same reasoning, the time and effort voters are willing to spend in deciding how to vote should ordinarily be small, since this time expenditure must compete with many private goods, the purchase and use of which also requires the individual's time. In contrast to consumers who can be expected to acquire an efficient amount of information about these private goods that they choose for themselves, we should not be surprised to find voters frequently leaving the difficult task of choosing alternative policies, a public good, in the hands of governments by deciding how to vote in a way that is at least as easy as the journey to the polls.

The question is, How does an electorate that is for good reason poorly informed about policies make effective its unsophisticated collective demand for "more utility"? How do we explain the fact that, unlike relatively well-informed consumers, "ignorant" voters regularly benefit greatly from decisions that they have participated in not at all? Notwithstanding our more idealized notion of active participatory democracy, the public-good aspect of voting should redirect our attention from the difficulties of forecasting prospective untested policies to the implication for electoral influence of retrospective reactions to actual policy chang This message is consistent with empirical studies of voter behavior wh display a marked voter tendency to avert abstract reasoning by deali in the concrete (Campbell et al. 1960, chaps. 9–10).[3]

II. Policy Influence through Performance Rating

To explain why the electorate's influence over policy choice is much broader than its participation in policy choice, we must clearly distinguish between the prospective and retrospective aspects of elections. The electorate may (1) *prospectively select the next government* by predicting which candidates and party will produce the greater utility in the next term of office,[4] and (2) *retrospectively approve or disapprove the incumbent government* at the end of its term according to its performance. Policymaking in term-of-office X is kept responsive by the anticipated competitive elections that are to follow X, not by the election that is over with prior to X. Once a government has been chosen and installed, it is just the politicians' need for voter approval at the end of their term that makes it costly for them to risk their careers (or reputations, as with last-term politicians) by alienating voters (Rothenberg 1965; Schlesinger 1966). Thus, though bygones are forever bygones, the fundamental fact remains that an electorate looking toward the future that failed to vote its past dissatisfactions would lose control over its political future. Retrospective evaluation is necessary for the control of government by electorates.

---

[3] Compare this paragraph with Buchanan (1970, pp. 113–17).

[4] As the voter compares the alternatives, he may consider: (1) his nonpolicy preferences such as party or other loyalties, personalities, etc. (just as a restaurant's utility depends on appearance as well as food); (2) the respective skills and competence of the competitors; (3) different policy goals and instruments that he expects the competitors to pursue if elected. Among these three aspects, it is only the extent to which the electorate chooses policy goals when it selects candidates and parties that we wish to question.

Conversely, though the ability of voters to *select* a government depends upon their ability to predict, it does not follow that the electorate must predict in order to *influence* incumbents. Electorates can passively influence specific government decisions without participating in these policy decisions in any way, without knowing which policies a potential government will favor, and without in fact ever understanding what the alternatives were. To prove this basic point, consider a group of unsophisticated voters who economize their political efforts. How can these voters collectively influence certain important policy decisions that will determine the trade-offs between employment and price stability, environmental amenities and industrial production, crime rates and tax burden, etc.? Their utility will be significantly affected by these trade-offs, but they do not understand them. They therefore lack ex ante preferences because they fail to predict what ranking of these policy goals will promote their utility. Further, they enjoy no pressure group representation on these issues. I assume government values the support of these voters.[5]

We must specify a vote criterion such that the support that the government will retain in the next election depends to a significant extent upon the utility consequences for voters of these kinds of policy decisions. Voters require a performance standard. *But they cannot derive this standard from the policy goals the opposition would have established* because, even if the opposition could be relied upon to reveal true policy intentions, we cannot expect unsophisticated voters to predict what utility difference the opposition's untested alternative policies would have produced. Thus voters cannot greatly economize their information costs by comparing actual policies with alternative policies in retrospect. They can evaluate the government without predicting only to the extent that they put aside hypothetical alternative policies by deriving their standards from their past experience.

---

[5] In the argument to follow we need *not* assume voter support is a maximand. The degree to which the personal preferences of governments are suppressed depends on whether most successful (office-holding) politicians are (1) maximizers (either by necessity or choice) or (2) "satisficers" who sacrifice both degrees of political security and the satisfaction of bigger pluralities to get a degree (what degree?) of discretionary freedom.

Accordingly, suppose at the end of a government's term each of our voters compares (1) that part of his utility for which he believes this government is responsible with (2) a personal performance standard that is what he has come to expect of the government. Each votes to support the incumbents only if either the level or the changes in this part of his utility satisfy his performance standard. Thus our voter drastically simplifies his decision in two ways: first, by using a performance standard; second, by holding responsible incumbents in general, the most prominent incumbent party, or the incumbent coalition. Whether he is thinking of collectively punishing (rewarding) this government for the past or using the past to estimate its future performance we need not bother to speculate upon.

Since each performance standard is personal, it varies from voter to voter. The crudely constructed standards of some voters exceed the utility that governments are capable of delivering; the standards of other voters understate what governments can accomplish. In effect, each voter's standard is his own minimum supply "price" for supporting the incumbents. Thus, when we order voters by their various supply prices, we obtain a rising supply schedule of voter support. The government tries to buy support with its policy resources during its term in order to reduce the risk of rejection at the end of its term. The amount of support it succeeds in acquiring varies continuously with the effect of its p decisions on the utility of voters who may in fact know nothing specific alternatives but who simply expect a certain level of perforn I conclude that passive unorganized voters who consider only thei experience, and thus attempt no policy predictions, can make it interest of governments to promote their utility without nec understanding how governments do it.

In order to cause a government to work for high performance ratings, these ratings must be *one* of the important determinants of shifts in the support of "old" voters and the division among new voters which will together decide its reelection or defeat. Though voter influence is necessarily related to government selection in this way, it is nonetheless essential that we do not equate influence with selection. Suppose the strength of the opposition and the rhetoric of the next election campaign were to outweigh performance ratings in deciding elections.[6] This would be compatible with the present theory of broad voter influence, since *it is not the theory's aim to explain who wins elections*; rather, this theory connects voter influence to the fact that incumbents, before they know the strategies of their next opponents, work to keep intact a foundation of support in voter satisfactions within a changing social system that is exceedingly difficult to control.[7]

---

[6] Actually, a number of studies have concluded that vote decisions are substantially set prior to the campaign. See Lazarsfeld, Berelson, and Gaudet (1944) and Key (1966) on presidential elections and Kramer (1971) on congressional elections.

The influence the electorate does assert through performance ratings depends on (1) what voters expect from government and (2) the government's ability to predict how policy alternatives will affect voters.

1. *Performance Pressures*

In progressive societies a rising level of voter expectations pushes governments to develop their policy capabilities. True, the performance standards of some groups are easy to satisfy, due either to their profound ignorance of the simple fact that government policies do affect them, or to their rigid ideology. Given their political preferences, these voters (and nonvoters) either react primarily to the personalities and life styles of conspicuous leaders, or are party loyalists, or are dedicated to whatever government is in office because they would prefer an aristocracy that judges its own actions to an accountable government that is finally judged by voters. But short of such extreme ignorance or ideology, it need *not* follow that the voter who knows less about a government's policies usually expects less from this government. Groups whose perceptions run mainly to their own concrete experience may expect *more* just because they do not fully appreciate how much a government is constrained by conflicting group demands; the policies of other domestic, foreign, and preceding governments; and many nongovernmental forces. Though a particular government will not respond to impossible demands that are beyond its capabilities, the pressure is on it to respond at the margin—that is, to favor policies that may add more support than they lose.

Moreover, this theory suggests that retrospective performance rating makes it easier for less informed voters to keep track of government. How, for example, does a government conceal the cost of regulations such as pollution abatement from voters who "naively" hold this government responsible for changes in their purchasing power? Voters who deal in ex post utility changes may indeed be said to "demand" specific public goods *as though* they were weighing marginal costs. Critics note that initial ex ante public opinion commonly supports a new policy such as foreign military intervention or price controls without understanding it, but the ex post opinion of voters who have measured the results of these policies against their expectations is probably much less subject to governmental (or media) manipulation chiefly because the performance standards of voters permit them to bypass the ambiguities of hypothetical policy goals and unstated costs (Cell 1973). Thus performance ratings indicate why voters sometimes blame governments for adversities though they cannot themselves explain what policy alternatives would have been more successful. We can also see why voters who apply performance ratings may be unaware of their collective influence over many specific policies.

---

[7] Explaining which party (or candidate) wins has been perhaps the chief objective of empirical work on elections. One important contribution that did not deal in ex ante preferences (Kramer 1971) found a significant correlation in this century between voter support of the president's congressional party and changes in pretax purchasing power. This statistical measure of performance omits nonmonetary aspects of real income (including changes in government services) and after-tax distributive changes but does reflect macro developments. Also see Stigler's (1973) critique and Okun's (1973) comments.

2. *The Knowledge of Governments*

Professional policymakers as a group, for all of their limited information and mistakes, have *far greater incentive* to discover how the alternatives would affect voters than do voters themselves. Successful politicians are, after all, professional dealers in policies whose careers depend upon both their political expertise in what will please voters and their technical expertise in how to produce the results that voters desire. The *comparative* knowledge of politicians and voters is basic to the interpretation of elections, as I now hope to show.

III. Some Implications for Elections

1. *Unreliability among Candidates*

I have argued that to keep policymaking accountable, v [illegible] evaluate performance retrospectively. Now I can specify tw [illegible] retrospective evaluation that produce accountability: (1) to the [illegible] that voters rate government by its utility consequences, they delegate policy evaluation to government and assert their demand for ex post performance, as explained in this paper; (2) to the extent that voters *also* rate government by its reliability in carrying through the specific policies they expected when they elected it, voters assert these ex ante policy preferences.

Several important implications follow from this generalized theory of accountability. If the electorate wishes to participate deliberately in policy choice when it votes, it must be able to predict what policy alternatives the parties and candidates embody. This is why participatory democracy requires that the electorate promote reliability by withdrawing support when a government fails to adhere to its promises (proposition 2).

But *accountability does not require reliability.* Regardless of what candidates and parties said they would do if elected, and whether these preelection statements reliably reveal their intentions, they will be judged by the utility consequences of what they in fact did in office (proposition 1).[8] Hence the incumbent's need for an ex post "yes" and the voter's willingness to register a simple ex post "no" surely extend the electorate's influence well beyond what policy preference voting alone would permit.

Furthermore, the promotion of reliability entails costs as well as benefits. Voters value reliability when they (1) think they know better than government (as indeed sometimes they doubtless do) which policy alternative is best, (2) simply want to participate, and (3) value candor in public life in itself apart from its utilitarian control value.

On the other hand, a strong preference for reliability may be costly when an electorate knows less about what policies it wants than does government. Suppose the results of an unreliable government's policies satisfy a solid majority by election time. How should a satisfied voter rank reliability as against demonstrated ex post competence on his two-value performance scale? Apparently, a strong preference for reliability would sometimes put competent policymakers out of office. Alternatively, suppose a government proves its competence in other respects, but adheres to one important policy promise which voters initially "approved" but later decide was a serious mistake because of costs far beyond what they had anticipated. Should voters return this reliable government to office despite this big mistake?[9] The more voters assume responsibility for the mistakes of instructed governments by rewarding ill-fated reliability, the more are governments encouraged to defer to uninformed voter opinion and the less is their incentive both to develop and to apply their leadership capabilities. Thus when successful candidates implement the promises they had to make to appeal to a poorly informed electorate, this reliability is not a free good because it reduces their freedom to promote voter satisfaction in office.[10]

Since reliability may be costly and is of limited value to an electorate that rates ex post performance, I do not expect politicians to be consistently reliable. I rather predict that (1) the electorate will rank performance over candor both by approving the successes of seemingly unpredictable governments and by punishing governments for bad policies with which voters themselves had initially agreed; and (2) campaign speech making will play a rather limited role in the development of democratic policies.

---

[8] The view that *reliability* as well as *communication* from voters and *divergence* among candidates and parties are all necessary conditions for voter influence is found in Downs (1957) and recently in Sullivan and O'Connor (1972).

[9] For example, say President Johnson had forthrightly stated his contingency escalation plans for Vietnam in the 1964 election and won. Should the many voters who would have approved these plans in 1964, because they were unable to predict the adverse consequences, have supported this administration in 1968?

[10] Consider the cost to voters if they had held Roosevelt to his 1932 balanced-budget pledge.

2. *Inefficient Communication*

Since voters can influence through performance ratings without communicating policy preferences, communication is surely at best a secondary function of elections. Elections are but one among a number of channels through which politicians inform themselves about voter utility functions. Election returns reflect a rich assortment of (1) ex post reactions to performance; (2) assessments of governing abilities, personalities, parties, etc.; and (3) ex ante policy opinions. The latter are of limited use to policymakers chiefly because they may be flaccid impressions rather than firm preferences. Further, to the extent that these opinions are instructive, elections are probably less efficient than opinion polls and constituency fieldwork in discovering them, both because candidate positions may be unclear to voters and because opinions are difficult to distill from the other electoral messages and noise.

3. *Convergence*

Despite their political expertise, professional politicians still disagree about how alternative programs will affect voter support. But as they become more knowledgeable about how to please voters, in their struggle
to avoid costly mistakes, should we not expect the competitors
becoming a more homogeneous (though competitive) profes
True, the more they converge, the less the opportunity for
to participate in policy choice through elections. But as th
converge on a policy issue because they know more, the les
for the electorate to intervene in this policy decision. C
less the knowledge of the competitors, the more they will t
increasing both the opportunity and comparative advantage
intervention.

Both because information is limited and because the nomination process casts up candidates of unequal political ability, convergence will remain incomplete, and it will be useful for voters to "judge the man." But imagine a world in which competing politicians and parties did become completely homogeneous in the policies they embodied (and were therefore unencumbered in recognizing and correcting their own mistakes in office); this would by no means render elections meaningless. Such homogeneity would eliminate the selection-of-a-government aspect of elections, even more fully delegate policy choice to governments, and further enhance the role of performance ratings with which voters keep policymakers accountable.

Aside from the public-good aspect of policy information, the following well-known propositions about the search for lower prices by rational consumers should also apply to the search for policy distinctions by voters.

The more prices differ, the greater the potential rewards of search. But price convergence among competitive products denotes a more perfectly informed market, not the absence of consumer influence. Moreover, search by sellers for consumers would promote convergence even though consumers were passive and uninformed (Stigler 1961, pp. 213–35). So, to conclude, the more the need to keep voters satisfied drives politicians to acquire information about voters, the less voters need to know about the intentions of politicians. Thus more fundamental than the outcome of particular elections are the underlying continuities as successive governments search for those policies that will secure ex post voter approval.

References

Becker, G. "Competition and Democracy." *J. Law Econ.* 1 (October 1958): 105–9.

———. *Economic Theory.* New York: Knopf, 1971.

Buchanan, J. M. *The Public Finances.* 3d ed., rev. Homewood, Ill.: Irwin, 1970.

Campbell, A.; Converse, P. E.; Miller, W. E.; and Stokes, D. E. *The American Voter.* New York: Wiley, 1960.

Cell, D. C. "A Theory of Voter Influence over Foreign Policy." In *Democratic Representation and Apportionment: Quantitative Methods, Measures, and Criteria,* edited by L. Papayanopoulos. *Annals of the New York Academy of Sciences* 219 (November 1973): 40–49.

Downs, A. *An Economic Theory of Democracy.* New York: Harper, 1957.

Due, J. F., and Friedlaender, A. F. *Government Finance.* 5th ed. Homewood, Ill.: Irwin, 1973.

Key, V. O., Jr. *The Responsible Electorate: Rationality in Presidential Voting, 1936–1960.* Cambridge, Mass.: Harvard Univ. Press, 1966.

Kramer, G. H. "Short-Term Fluctuations in U.S. Voting Behavior, 1896–1964." *American Polit. Sci. Rev.* 65, no. 1 (March 1971): 131–43.

Lazarsfeld, P. F.; Berelson, B.; and Gaudet, H. *The People's Choice.* New York: Duell, Sloan, & Pearce, 1944.

Okun, A. M. "Comments on Stigler's Paper." *A.E.R.* 63, no. 2 (May 1973): 172–77.

Olson, M., Jr. *The Logic of Collective Action.* Cambridge, Mass.: Harvard Univ. Press, 1965.

Rothenberg, J. "A Model of Economic and Political Decision Making." In *The Public Economy of Urban Communities,* edited by J. Margolis. Baltimore: Johns Hopkins Press, 1965.

Schlesinger, J. A. *Ambition and Politics: Political Careers in the United States.* Chicago: Rand McNally, 1966.

Schumpeter, J. A. *Capitalism, Socialism, and Democracy.* 3d ed. New York: Harper, 1950.

Steiner, P. O. "The Public Sector and the Public Interest." In *Public Expenditures and Policy Analysis,* edited by R. H. Haveman and J. Margolis. Chicago: Markham, 1970.

Stigler, G. J. "The Economics of Information." *J.P.E.* 69, no. 3 (June 1961): 213–25.

———. "General Economic Conditions and National Elections." *A.E.R.* 63, no. 2 (May 1973): 160–67.

Sullivan, J. L., and O'Connor, R. E. "Electoral Choice and Popular Control of Public Policy." *American Polit. Sci. Rev.* 66, no. 4 (December 1972): 1256–68.

Tulloch, G. *Toward a Mathematics of Politics.* Ann Arbor: Univ. Michigan Press, 1967.

Time-Series Changes in Personal Income Inequality in the United States from 1939 with Projections to 1985: A Comment

Aba Schwartz and Donald O. Parsons

Ohio State University

In a recent article Chiswick and Mincer (1972) attempt to explain income inequality by using a human capital theoretic approach. They measure income inequality by the variance of the log of net earnings. As they argue, human capital considerations imply a dependence of net earnings in a predictable way on number of years of schooling, number of years of experience on the job, number of weeks worked throughout the year, and *rates of return on self-investment made throughout the life span.* The variance of the log of net earnings can then be expressed in terms of the variances and covariances of the explanatory variables (some en a product form). With this implied relationship one can co implied variance of log net earnings using data on the distribu explanatory variables. This variance is then compared with th variance of log net earnings. It turns out that the human capi capable of explaining a remarkable 69 percent of this observe This research must certainly be judged a bold and heroic enterp

Unfortunately, as we intend to demonstrate below, the "theoretic model which Chiswick and Mincer develop and then apply empirically corresponds only in the most superficial way to their own assumptions or a truly capital theoretic approach. In what follows we first sketch out the basic Chiswick-Mincer (CM) model and then contrast this with a rigorously developed human capital model based on their assumptions. Particular attention is paid to the innovative elements in the model, namely, measurement of years of experience and the role of annual weeks worked. The results indicate that the variance (and expectation) expression of log earnings is completely misspecified. The "true" model apparently is too complex to be useful for income inequality studies, even under the strong and frequently unintuitive assumptions of Chiswick and Mincer.

The authors are visiting associate professor and associate professor, respectively, at the Ohio State University.

[*Journal of Political Economy*, 1974, vol. 82, no. 5]

© 1974 by The University of Chicago. All rights reserved.

The authors start with the basic annual gross earnings function for an individual,

$$E_j = E_o + \sum_{t=1}^{j-1} r_t C_t, \tag{1}$$

where E_j = gross earnings in year j, E_o. = annual earnings due to original endowment, C_t = value of human capital investment in year t, and r_t = rate of return to investment in year t. Defining the proportion of gross earnings in year t which is self-invested as k_t, namely,

$$k_t = \frac{C_t}{E_t}, \tag{2}$$

equation (1) can be transformed into

$$E_j = E_o + \sum_{t=1}^{j-1} r_t k_t E_t. \tag{3}$$

Equation (3) is recursive in the form

$$E_j = E_{j-1} + r_{j-1} k_{j-1} E_{j-1} \tag{4}$$

with solution

$$E_j = E_o \prod_{t=1}^{j-1} (1 + r_t k_t). \tag{5}$$

These are all standard manipulations, and the confusion begins when the authors explicitly introduce the effect of number of weeks worked (to be denoted W) on the gross earnings function. On the basis of an empirical study by Mincer (1972), who estimates the elasticity of annual earnings with respect to number of weeks worked (γ) to be 1.2, the authors alter equation (5) into

$$E_j = W_t^{\gamma} E_o \prod_{t=1}^{j-1} (1 + r_t k_t). \tag{6}$$

Clearly this requires a redefinition of E_o which is never given, but that is a detail. Equation (6) is approximated in logarithmic form with assumptions that during the schooling period S the rate of return is constant at r and $k = 1$, so

$$\ln E_j = \ln E_o + rS + \sum_{t=S+1}^{j-1} r_t k_t + \gamma \ln W_j. \tag{6a}$$

This essentially is the model they develop empirical counterparts for, before finding and exploiting the variance expression.

Consider for a moment the nature of the introduction of number of weeks worked. The authors recognize at several points the suspected link between average work week and specific human capital stocks (see pp. 37, 40), which are presumably correlated with their measure of experience, but the authors indicate no desire to integrate this functional relationship into their model. Even if one ignores this problem, a little reflection suggests some caution in introducing weeks worked in a haphazard way. Both rate of return on human capital stock and investment costs will be influenced by weeks worked per year. For a given productivity, clearly a reduction in weeks worked will reduce the annual rate of return expected on a given capital stock. Similarly fewer work hours, if demand determined, imply lower opportunity costs of investment. In this model, where investment is some fraction of weeks worked (through gross earnings), lower work weeks imply less investment.[1]

Returning to the basic equation (3), the introduction of weeks worked into the calculations suggests that the rate of return must be indexed by every return period as well as by the cost period, since the rate of return will vary with weeks worked. Likewise, E_o should be indexed by j, as it also depends upon the number of weeks worked;

$$E_j = E_{oj} + \sum_{t=1}^{j-1} r_{tj} k_t E_t, \tag{7}$$

where r_{tj} = rate of return in period j on investments in human capital undertaken in period t, and E_{oj} = return in year j due to inherited capabilities.

Equation (7) can be equivalently expressed as

$$E_j = E_{j-1} + r_{j-1,j} k_{j-1} E_{j-1} + E_{oj} - E_{oj-1} + \sum_{t=1}^{j-2} (r_{tj} -$$

This simply says that earnings in this period are equal to ca period plus the relevant returns on any last-period investment necessary corrections in rate of return between the two periods that mig result from different weeks worked.

If we assume (as CM do) that the elasticity of annual earnings with respect to weeks worked is constant, it can be shown that the following relations hold:[2]

$$r_{tj}(W_j) = r_t^*(1) W_j^\gamma; \tag{9a}$$

$$E_{oj}(W_j) = E_o^*(1) W_j^\gamma; \tag{9b}$$

$$E_t(W_t) = E_t^*(1) W_t^\gamma, \tag{9c}$$

where $r_t^*(1)$ = the rate of return per week of investment made at year t, *if only 1 week is worked,* $E_o^*(1)$ = the weekly earnings due to inherited capability *if only 1 week is worked*, $E_t^*(1)$ = the weekly earnings in year *t if only 1 week is worked*, and γ = the elasticity of gross annual earnings with respect to number of weeks worked. Substituting (9) in (8) and rearranging terms yield

---

[1] Certainly k itself will be a function of W. Since total training time will vary directly with W, holding k constant, it is unclear what direction the proportion of time devoted to training will take. See Lindsay (1971) for an interesting discussion of the interrelationship of work hours desired and human capital investment decisions.

[2] This assumption made by CM is based on empirical findings by Mincer (1972).

$$E_j(W_j) = E_{j-1}(W_j) + r_{j-1}^*(1) W_{j-1}^\gamma k_{j-1} E_{j-1}(W_j). \qquad (10)$$

Equation (10) is a first-difference equation which, when solved recursively, yields

$$E_j(W_j) = E_o(W_j) \prod_{t=1}^{j-1} (1 + r_t^*(1) W_t^\gamma k_t). \qquad (11)$$

Clearly equation (11) cannot be simplified in any way to resemble equation (6a). So we assume (as CM do) that:

a) $W_t = W_{t+i} = W$ for all t and i. Under this assumption the last three elements in (8) vanish, and correspondingly the weeks-worked variable (W) in equation (11) is not time subscripted.

b) There is a constant weekly rate of return,[3] say $r_s^*(1)$, over the entire schooling period S.

c) There is a constant weekly rate of return, say $r_e^*(1)$, over the entire "on-the-job training" period.

d) $k_t = 1$ for the entire schooling period.

Imposing these assumptions and transforming (11) into logs yields

$$\begin{aligned} \ln E_j(W) &= \ln E_o(W) + \ln (1 + r_s^*(1) W^\gamma) S \\ &\quad + \sum_{t=S+1}^{j-1} \ln (1 + r_e^*(1) W^\gamma k_t) \qquad (12) \\ &\cong \ln E_o(W) + r_s^*(1) W^\gamma S + r_e^*(1) W^\gamma \cdot \sum_{t=S+1}^{j-1} k_t, \end{aligned}$$

which is the human capital counterpart of CM's equation (6a) above. Notice that the schooling and on-the-job training expressions must be corrected for the individual's weeks worked.

Net earnings y_j, that part of gross earnings in period j not invested, is the empirically more useful concept and can be represented as simply

$$y_j = (1 - k_j) E_j, \qquad (13)$$

or in logarithmic form,

$$\ln y_j = \ln (1 - k_j) + \ln E_o(W) + r_s^*(1) W^\gamma S + r_e^*(1) W^\gamma \sum_{t=S+1}^{j-1} k_t \qquad (14)$$

for $j > S$.

---

[3] **Actually CM assume a constant *annual* rate of return, but in conjunction with constant *W* the two assumptions are equivalent.**

CM's assumption is that

$$k_t = k_o\left(1 - \frac{T_t}{T^*}\right), \tag{15}$$

where k_o is a constant, T_t is years of experience in year t, and T^* is the terminal number of years of investment in on-the-job training. This leads CM to the approximation:

$$\sum k_t = k_o T - \frac{k_o}{2T^*} T^2, \tag{16}$$

where T is the number of current years of experience. They also approximate ln $(1 - k_j)$, where k_j is defined by (15), by a Taylor expansion truncated after the third term. This approximation yields a polynomial of second degree in T. Where these two approximations are introduced into (14), we end finally with a linear equation in the variables $W^\gamma S$, $W^\gamma T$, $W^\gamma T^2$, T, and T^2. This bears no simple relationship to the CM form of the log earnings and leads to a different variance expression even when one assumes that none of these terms are correlated with the weekly rate of return coefficients, $r_s^*(1)$ and $r_e^*(1)$, and that these last two terms are uncorrelated with each other as CM posit.[4]

Of course, it is possible that CM may have captured some element of empirical truth in their earnings function even though it is mis A brief examination of some simple interrelationships suggests that this is not true. Consider the indirect implications of CM' the quantity of on-the-job training undertaken. CM state (do not prove that the coefficient of experience r' is approxim

$$r' \cong \frac{r_e k_o}{2},$$

where r_e is the annual rate of return on on-the-job training. If one accepts CM's estimate of $r' = 0.04$ and the roughly established fact that $r_e = r_s$ (see Mincer [1962]), the rate of return on schooling where $r_s = 0.11$, an estimate of the quantity of time devoted to investment in the first postschooling year (k_o) can be derived. In this case,

$$k_o = \frac{2r'}{r_e} = \frac{2(0.04)}{(0.11)} = 0.73.$$

---

[4] Even if we assume that $r_s^*(1)$ is uncorrelated with $r_e^*(1)$ and each is uncorrelated with W (across individuals), then

$$\mathrm{Cov}[r_s(W), r_e(W)] = E[r_s^*(1)]E[r_e^*(1)]\mathrm{Var}(W^\gamma),$$

where E denotes expected value and Var denotes variance. It follows that CM's assumption (p. 539) that the *annual rate of return* to schooling is uncorrelated with the *annual rate of return* to experience is unacceptable.

It is intuitively very hard to accept that the average individual just out of school, were he not to invest, could earn income about four times what he actually earns. Add to this estimate the assumption implicit in the CM model that individuals continue to invest until they leave the labor market (since eq. [15], and therefore CM's earnings function, is only applicable as long as $T \leq T^*$, which they nonetheless use for all ages), and you can derive an estimate of total full-time equivalent years devoted to on-the-job training, namely,

$$\text{Full time years} = \frac{k_o(T^* + 1)}{2} = 0.365\ [T^* + 1].$$

If $T^* = 40$ years, this model estimates that out of 40 working years, 15.0 or roughly 36 percent will be devoted full time to investment. Clearly some parts of the CM model are radically incorrect, and one need go back only to their original misspecification of the model to find a reasonable reason why.

References

Chiswick, Barry R., and Mincer, Jacob. "Time-Series Changes in Personal Income Inequality in the United States from 1939, with Projections to 1985." *J.P.E.* 80 (suppl.; May/June 1972): 534–66.

Lindsay, C. M. "Measuring Human Capital Returns." *J.P.E.* 79 (November/December 1971): 1195–1215.

Mincer, Jacob. "On-the-Job Training: Costs, Returns, and Some Implications." *J.P.E.* 70 (suppl.; October 1962): 50–79.

———. "Schooling, Age, and Earnings." In *Human Capital and Personal Income Distribution.* New York: Nat. Bur. Econ. Res., 1972.

Time-Series Changes in Personal Income Inequality in the United States from 1939 with Projections to 1985: Reply

Barry Chiswick and Jacob Mincer

Columbia University

Our commentators (henceforth P-S) cannot be faulted for pointing out a fact of which we are fully aware: the inclusion of the "weeks worked" variable in our model was not rigorous in the sense of explicitly formulating a dependence on human capital, labor supply, and other considerations. We preferred to leave this part of the story out for the time being. Work on this subject is only beginning at the NBER and elsewhere. The problems posed by P-S should be of value in such work.

The specific functional form for the inclusion of weeks worked was adopted to facilitate the empirical application of the model. We
the insertion of W in a somewhat careless way in the beginning
on our page 37. The sentence should read: "If E^* is full-ye
actual annual earnings are $E_t = E_t^* W_t^\gamma$, where W is the fracti
worked during the year." Thus, r^* is the full-time rate of ret
is defined as C_t/E_t^*. At this level we do not see any particul
arising with this formulation. The first sentence of the first p
on page 37 says, in effect, that the appropriate variables and param
(including r and E_0) should be starred. Of course, there is no star on the left-hand side of our equation (4).

It is calculated by P-S that our parameter estimates imply $k_0 = 0.73$, which is rather implausible. Actually, the appropriate coefficient of T is not r^*k as they suggest, but $r^*k_0 + k_0/T^*(1 + k_0)$, as we show on page 38. The T^* is the period of positive net investment, which is not assumed but implicit in the income profile. Its value is about 20 years. If so, $k_0 = 0.44$, and total "post-school year equivalents" of positive net investment $K = 4.4$, not 15 as P-S estimate. These estimates are not very secure, but not clearly outrageous. We should keep in mind that these as well as the "rate of return" estimates are not the main targets, but merely by-products of a much-simplified model whose purpose was to study time changes in the aggregate-income variance.

[*Journal of Political Economy*, 1974, vol. 82, no. 5]
© 1974 by The University of Chicago. All rights reserved.

We are puzzled by the P-S view that the "true" model would be "too complex to be useful for income inequality studies." Though our initial effort is far from being comprehensive, we clearly prefer a useful simplification to a useless complexity.

Market Allocation of Exhaustive Resources

Oliver Scott Goldsmith

University of Wisconsin

The present concern for natural resource conservation has resulted in a surge of interest in the economic theory of the mine. First introduced by Gray (1914), this analysis has generally centered upon the optimal time stream of output for an individual firm in an industry of mines. It was only recently that Richard Gordon attempted to characterize the competitive-market solution for an industry.[1] His result, that a purely competitive-market solution need not be efficient, is odd in that the inefficiency does not arise from any externality or nonconvexity. His technique was to derive the production plan of an individual mine under the assumption that the mine manager maximizes the present value of his resource. This solution was contrasted with that of a socialist manager who maximizes the present value of the resource of the whole industry while acting as a price taker. This note will show that under t assumptions, the purely competitive solution is indeed equivale of the price-taking socialist manager. Furthermore, both solu economically efficient.

The theory of the mine is an instrument for determining the opt put pattern over time for a fixed-stock resource, assuming that the will be entirely used. Assuming with Gordon that the total stock resource is known and of homogeneous quality, and also that production from any mine in any time period does not affect production costs in any later period, the single-mine firm will maximize the present value of the expression

$$\int_{t_1}^{t_2} \pi[q(t), t]e^{-rt}\, dt - \lambda\left[\int_{t_1}^{t_2} q(t)\, dt - K\right],$$

where π denotes the profit function in which $q(t)$, the quantity produced at time t, and time itself serve as arguments. The mining firm's rate of discount is r, and the fixed total stock of the resource it controls is K. Solution of this expression yields the dates for beginning and terminating

---

The author gratefully acknowledges the guidance of Donald A. Nichols in writing this paper. Responsibility for all errors of course lies with the author.

[1] Gordon (1967).

[*Journal of Political Economy*, 1974, vol. 82, no. 5]

© 1974 by The University of Chicago. All rights reserved.

production as well as the Euler equation, which determines the pattern of production between the beginning and terminating dates. The latter can be simplified to

$$M\pi(t) = MR(t) - MC(t) = \lambda e^{rt},$$

where $M\pi$ is marginal profit, MR is marginal revenue, and MC is marginal production cost. The Lagrangian multiplier λ is usually called a user cost and reflects the fact that whenever the total stock of resources is limited, greater sales today imply fewer sales in the future. This rule tells the firm to allocate its production in such a way that the discounted marginal profit is the same at any instant in time. Thus, marginal profit rises at the rate r and marginal revenue exceeds marginal production cost at every instant of time. The terminal condition requires that at t_2 marginal profit equal average profit; that is, the competitive firm must then operate at the point of minimum average cost.

To this point the analysis is straightforward; confusion arises when Gordon examines the competitive industry. Is the efficient-firm behavior also efficient for the industry? Let us, like Gordon, derive the solution for a socialist manager who maximizes the present value of the total resource while acting as a price taker and see if the decision rules differ from those of an industry of purely competitive mining firms.

The socialist would seek to maximize the present value of the expression

$$\begin{aligned} &\int_0^T W[q_1(t), q_2(t), \ldots, q_n(t), t]e^{-rt}\,dt \\ &-\lambda_1 \int_{{}_1t_1}^{{}_1t_2} q_1(t)\,dt - K_1 \\ &\qquad\vdots \\ &-\lambda_n \int_{{}_nt_1}^{{}_nt_2} q_n(t)\,dt - K_n, \end{aligned}$$

where the subscripts now indicate the individual mines such that

$$\sum^n q_i(t) = \text{total industry production at } t,$$

$$\sum^n K_i = \text{total industry resource stock.}$$

The integral on the objective function runs over the period of production for the industry as a whole, while the integrals on the constraints each run over the period of production for the particular mine only. The W indicates a welfare function whereby it is assumed that the socialist can maximize utilities and allocate resources efficiently by being a price taker.

The solution to this problem is characterized by a set of $2n$ terminal conditions that give the time intervals for production from each particular mine,[2] by the equality of the n resource stock constraints, and by a set of n Euler equations of the following form

$$[P(t) - MC_i(q_i)]e^{-rt} = \lambda_i,$$

where P is the output price, which is unaffected by the output of any individual mine and MC_i is marginal production cost of the ith firm, which is a function only of output of the ith firm. The λ_i is the user cost associated with the ith mine and depends upon the cost function of the mine, its stock of reserves, and on total supply and demand for the industry's output. For example, a mine with low marginal costs and a small stock of reserves would have a large λ_i value relative to a mine with high marginal cost and large reserves. In any period the marginal production costs among mines need not be equalized.

It is in his characterization of the efficient socialist solution that Gordon errs, for he expects the socialist to "minimize production costs in a given period by equating marginal production costs among operating firms."[3] In fact marginal production costs will be equated in only two special cases—when costs are constant, or when all mines are identical in terms of resource stocks and cost curves. In general, the previous equations show that for each mine marginal production cost plus marginal user cost must equal price.

Intuitively, equalization of marginal production costs across
will lower current production costs by increasing output fo
efficient mines and reducing it for those less efficient. How
situation raises future production costs since the efficient mine
rapidly depleted, and production must be increased from the
mines. This cost is what is reflected in the λ_i and when add
marginal production cost gives the true cost of another unit of out
from any mine site. A simple numerical example in the Appendix confirms this.

Equalization of costs across mines occurs in the efficient socialist solution only in the sense of cost defined as the sum of direct production costs and user costs. Gordon does not let the socialist use all information available to him concerning differences among mines with respect to cost functions and resource stocks that is reflected in the differing user costs. Competitive firms would, of course, use this information, and the socialist would be a poor manager if he did not do the same. Gordon incorrectly implies that an industry-wide user cost exists that is the same for each mine.

The solution derived for each mine operated by a socialist manager is identical to the one that would result if the mine were operated by a purely competitive firm. The socialist requires

---

[2] Since at least one mine will operate in the first and last production periods, these $2n$ conditions determine the time interval of total production.

[3] Gordon (1967, p. 282).

$$[P(t) - MC_i(q_i)]e^{-rt} = \lambda_i$$

for each site, and pure competition would require the same result for any single-mine firm.

In each problem, supply and demand for the resource are the same, and each mine site is operating under the same cost function and quantity constraint. The perfectly competitive firm is unable to influence price by its production decision, while the socialist does not attempt to effect a high price, so neither formulation constrains the price of output that arises in the solution. Since that price is derived as a part of the solution to the problem, its pattern must be such that the stock of the resource is exactly exhausted at time T in each case. Thus, since each problem is the identical cost-minimization problem, the output stream determined for each mine, as well as the pattern of price over time, is identical in either case.

Furthermore, as is usual with pure competition, there is no incentive for merger, since the sum of profits from individual mines is the same whether their production schedules are individually or jointly determined. Only if firms become large enough so that price and marginal revenue are no longer equal can merger increase total profits.

The efficiency of the solution characterized above depends upon the nature of the welfare function that the socialist maximizes. If he maximizes his profits subject to the condition that he act as a price taker, the efficient solution will result. In that case, his behavior will be exactly analogous to that of the mining firm that faces a price it cannot affect and produces to the point at which total marginal cost, consisting of marginal production cost and marginal user cost, equals price.

Alternatively, efficiency of the solution can be seen by considering a socialist manager who must rent the n mines yearly from n different owners. The cost of rental of the mines becomes an integral part of the cost of production for the socialist, and he sets the marginal production cost, including mine rental, equal to price to get an efficient solution. The individual mine owner sees his mine as a stock, the present value of which he desires to maximize. The rent he can charge depends upon the costs of production and the amount of resource available at his particular mine site, as well as total supply and demand for the industry's product; at the maximum, it must rise at the rate r. That price at each site should be just equal to the λ_i that the socialist would calculate himself at each site if they were state-owned. Given private ownership, however, the λ_i represent production costs that must be paid, and in this case marginal production costs as viewed by the socialist are equal across all mines within each period.

The analysis easily allows for quality differences within the stock of the resource. Then each competitive-mine firm faces the same problem as the socialist manager if one defines the price of output in terms of a given quality of refined ore and defines a cost function for each grade of ore. The same implications follow in that each grade of ore will have a user cost within the firm, and marginal production cost across grades of ore will not be equalized in the profit-maximizing solution. Just as the socialist manager would extract ore simultaneously from mines with different marginal costs when only one grade of ore is present, the competitive-mine firm would extract different grades of ore simultaneously. Thus, in an efficient competitive or socialist industry low-grade ore could be mined at some sites prior to high-grade ore at others.

Thus, the purely competitive solution is efficient, and when Gordon's socialist equates marginal production costs within a period his production scheme will be inefficient.

Appendix

Counterexample to Gordon Rule for Socialists

Assume a very simple world in which the output from the only two mines may be sold in any quantity and in any time period for $5.00. The large mine, B, has a three-unit stock and the small mine, S, has only one unit, but the mines have identical upward-sloping cost curves characterized as follows:

Units of Output	Marginal Cost
1	$2
2	$3
3	$4

Four possible time schemes of production can be compared as to present value at a discount rate of 100 percent. The first three, which equate marginal cost between the mines in each period, are all dominated by the fourth, which implicitly allows user cost to affect the production stream.

	Units Produced by Period			Profit by Period			
	1	2	3	1	2	3	Present Value
Case 1:							
S	1	...	...	$3	...	...	
B	1	2	...	$3	$5	...	$8.50
Case 2:							
S	1	...	...	$3	...	...	
B	1	1	1	$3	$3	$3	$8.25
Case 3:							
S	...	1	...	...	$3	...	
B	2	1	...	$5	$3	...	$8.00
Case 4:							
S	1	...	...	$3	...	...	
B	2	1	...	$5	$3	...	$9.50

References

Gordon, Richard. "A Reinterpretation of the Pure Theory of Exhaustion." *J.P.E.* 75, no. 3 (June 1967): 274–86.

Gray, L. C. "Rent under the Assumption of Exhaustibility." *Q.J.E.* 28 (May 1914): 66–89.

How Analysis Affects Decision-Makers' Views: A Bayesian Model

Rolla Edward Park

RAND Corporation

The analysis of public policy issues is often a frustrating business. After preparing a careful study of the likely effects of a proposed policy, the analyst is apt to find that people in decision-making positions have neither the time nor the training to read it carefully. Decision makers may know who did the study and what the major results were, but they are unlikely to understand how (or how well) it was done. Nor are they likely to believe the results of any single study, no matter how carefully it was done.

In one recent case—the Federal Communications Commission's cable television rule making—the decision makers appeared to dismiss in a few sentences all of the analyses they had received:

> The conflicting conclusions of these studies make abu clear the difficulties involved in attempting to predict th when there are so many variables and unknowns. Wh reports and studies have been useful in illuminating the elements of our policy decision, we cannot rely on any part report or study as a sure barometer of the future. We would simply point out there is no consensus, and we do not pretend that we can now forecast precisely how cable will evolve in major markets. There is inherent uncertainty.[1]

Under these circumstances, can the analyst have any influence on the views held by decision makers? I argue in this note that he can. I use a Bayesian model that explicitly recognizes the "inherent uncertainty":

An earlier version of this paper was presented at a panel discussion held in August 1972 at the annual meeting of the Western Economic Association. The proceedings volume for the panel discussion, entitled *The Role of Analysis in Regulatory Decisionmaking: The Case of Cable Television*, is available from D. C. Heath Co., Lexington, Massachusetts. I appreciate helpful comments by John Rolph and the referee. The views expressed are mine alone, not those of the RAND Corporation or its research sponsors.

[1] FCC (1972), p. 3261.

[*Journal of Political Economy*, 1974, vol. 82, no. 5]

© 1974 by The University of Chicago. All rights reserved.

decision-makers' views both before and after policy analysis are probabilistic. Also, the model shows that decision makers do not have to read and understand policy analysis for the analysis to be effective.

To be concrete, I apply the model to an actual issue that received a good deal of analytical attention during the FCC's proceedings on cable television. One of the central issues in the proceedings was the impact that cable might have on over-the-air broadcasting. Cable carrying distant signals will fragment local stations' audiences and tend to reduce their advertising revenues. This may force some stations off the air and make others cut back on the quantity and quality of programming. Thus, people who do not subscribe to cable may have their television service reduced—a cost to society that must be balanced against the potential benefits of cable.

Of particular concern to the FCC was the possible impact of cable on nonnetwork (independent) UHF stations. For one thing, these stations are almost all in financial straits (only five out of 48 reported a profit in 1970) and so may be particularly likely to stop broadcasting if their revenues are reduced. For another, the FCC has a firm historical commitment to promote UHF broadcasting as the preferred method to increase the number of television signals available in any locality.

My concrete example, then, is the evolution of the FCC's views[2] about the likely impact of cable growth on independent UHF television stations. In particular, I apply the Bayesian model to show how the change in views can be explained as a result of analysis. For statements of views before and after analysis, I turn to the FCC's own language.

Changing FCC Views: Verbal Statement

At the opening of the cable television proceedings (docket no. 18397-A) in June 1970, the FCC was convinced that the threat of cable to UHF was real and serious, as this preanalysis quotation indicates:

> CATV operating with distant signals can achieve significant penetration figures in the major markets—most probably in the order of 50 percent. . . . We [are] convinced that a penetration of this order could pose a real threat to UHF development and that the unfair competition would be significant.[3]

During the next several months (July 1970–February 1971) a number of analyses of this subject were submitted to the FCC. An FCC staff study[4] pointed out that cable puts UHF stations on the same technical footing as VHF stations by overcoming a number of reception and tuning problems peculiar to UHF. This technical help would tend to offset the harm caused by audience fragmentation. Research done at RAND[5] concluded that, for UHF independents, the technical help would dominate audience fragmentation, so that cable would actually increase these stations' revenues by about 20 percent in the mid-1970s. Analyses filed by a number of interested parties (broadcasters)[6] criticized the methods and conclusions of the staff study and the RAND reports.

---

[2] In reality, of course, only commissioners have views, not commissions. "Commission's views" is just a convenient shorthand for "views stated in documents adopted by the commission"—presumably an indication of some sort of central tendency of the views of individual commissioners.

[3] FCC (1968). This view is referred to and reaffirmed in FCC (1970*b*).

[4] FCC (1970*a*).

By August 1971, the FCC's views as to the likely impact of cable on UHF had changed considerably from those held before exposure to the analyses mentioned above and related arguments, as this postanalysis quotation indicates:

> Broadcasters argue that any distant signal cable policy will have a disastrous impact on already shaky UHF stations. On the other hand, we have independent studies such as those submitted by the RAND Corporation suggesting that UHF will be likelier helped than hurt by cable—because UHF is still handicapped by reception problems, and these problems disappear with carriage on cable. Our own study of the matter has persuaded us that it would be wrong to halt cable dev ment on the basis of conjectures as to its impact stations. We believe the improvements that cable will clearer UHF coverage will at least offset the inroads audiences made by the limited number of distant sig our rules would permit to be carried.[7]

The Bayesian Framework

Did the analyses filed with the FCC bring about this change in views? This question is impossible to answer with certainty, but I argue that a yes answer is plausible. To do so, I "fit" a simple Bayesian model to explain the change in FCC views.

Bayesian statistical theory provides formal methods for making adjustments in probabilistic views as new evidence comes in. In the abstract, the Bayesian framework is as follows. Say you are interested in which of n mutually exclusive, exhaustive alternatives, A_i, $i = 1, \ldots, n$, is true. Even before performing an experiment designed to shed some light on the question, you have some idea about how probable the different alternatives are. These prior probabilities, $P(A_i \mid H)$, are based on the background information, H, that you have before you do the experiment.

---

[5] Park (1970, 1971).

[6] American Broadcasting Company (1970, 1971), Association of Maximum Service Telecasters (1970, 1971), Kaiser Broadcasting Corporation (1970), National Association of Broadcasters (1970, 1971), and Twenty-one Television Stations (1970, 1971).

[7] FCC (1971). A similar statement accompanies the rules finally adopted in FCC (1972).

You perform the experiment and find that the outcome is X. You know how probable X is if each of the alternatives is true. That is, you know $P(X \mid A_i, H)$. Then Bayes's theorem gives you a way to calculate posterior probabilities for the alternatives, that is, probabilities given the way the experiment turned out:

$$P(A_i \mid X, H) = \frac{P(A_i \mid H)\, P(X \mid A_i, H)}{\Sigma\, P(A_i \mid H)\, P(X \mid A_i, H)}.$$

In most conventional applications of Bayes's theorem, $P(A_i \mid H)$ is a subjective prior probability and $P(X \mid A_i, H)$ is a bona fide likelihood function in the frequency sense. In this note, in contrast, it is easiest to interpret all of the P's as subjective probabilities. In either case, the formula gives the only consistent way to calculate posterior probabilities.

Changing FCC Views: Numerical Probability Statement

To apply this framework to the evolution of the FCC's views on the impact of cable on UHF, it is necessary to express those views as numerical probabilities. This is unavoidably a subjective process. The pre- and post-analysis quotations above certainly do not unambiguously translate into a firm set of numbers, but I have tried to assign numerical probabilities that are a reasonable reflection of the FCC's expressed views.

For simplicity, I assume that the FCC's views were first modified after receipt of the staff study and the RAND reports, then modified again after receipt of the critical analyses done by the various interested parties. The actual chronology was much more complicated than that, but it does not matter. I also assume that the alternatives (A_i) are trichotomous. The three alternatives are: (1) helps: cable will significantly help independent UHF stations; (2) neutral: cable will not affect independent UHF stations very much one way or the other; and (3) hurts: cable will significantly hurt independent UHF stations.

Prior to receipt of any of the analyses filed in docket no. 18397-A, the FCC's views were as given in the preanalysis quotation above. A reasonable translation of these views into numerical probabilities for the three alternatives is shown in the first column of table 1. That is, the FCC was quite sure that cable would hurt UHF (P[hurts] = .7). The other alternatives could not be completely ruled out, though. For one thing, there was still some uncertainty about how many people would actually subscribe to cable in large cities with good over-the-air television service. For another thing, the FCC was already aware that cable improves

TABLE 1

CHANGE IN FCC VIEWS DUE TO STAFF STUDY AND RAND REPORTS

Alternative	Prior Probability	$P(X \mid \text{Alternative})$	Posterior Probability
Helps	.1	.9	.32
Neutral	.2	.6	.43
Hurts	.7	.1	.25

reception of UHF signals.[8] But the probabilities were thought to be small, say $P(\text{helps}) = .1$ and $P(\text{neutral}) = .2$.

The new observation, X, is the appearance of the staff study and RAND reports containing the views described above.[9] How did they affect the FCC's views? The question in this framework is not whether the FCC believed them or not but, rather, how probable the appearance of these reports was, if the various alternatives were true. The staff study did not unequivocally support any of the alternatives, so how you answer this question depends largely on what you think about the objectivity and competence of the RAND research.

A plausible representation of the FCC's assessment is shown in the second column of table 1.[10] If cable really helps UHF, there is a high probability that RAND would be able to figure that out and
report that says so ($P[X \mid \text{helps}] = .9$). Even if cable does not d
for UHF one way or the other, it is not too improbable that the
reports would say that it helped. This might be partly because th
dary between the two alternatives is fuzzy and a small miscal
could slide the conclusion one way or the other. Or it might be b
(as some have claimed) RAND has a procable bias. Or there may
bias in any research in favor of definite, interesting conclusions. So perhaps $P(X \mid \text{neutral}) = .6$. But if cable does actually significantly hurt UHF, it does not seem probable that RAND could be biased enough or incompetent enough to say it helps ($P[X \mid \text{hurts}] = .1$).

Applying Bayes's theorem, we get the posterior probabilities shown in the last column of table 1. Note that the FCC is less certain about the effect of cable on UHF after considering the staff study and the RAND reports than it was before. Before, the probability of "hurts" was quite

---

[8] See FCC (1970*b*), par. 3.

[9] In the model, X is dichotomous; either the staff and RAND conclude that cable will help UHF or they do not. The details of the reports do not matter. In reality, I believe that this aggregate approach is a fairly good representation of FCC perceptions; they certainly knew the major conclusions of the reports but were not concerned with details.

[10] Here the assignment of probabilities is even more subjective than it was for the priors, because there is not any published expression of FCC views to serve as a guide. I have had to rely on my own impressions as a participant in the proceedings, and I hope that the reader will agree that the numbers I assign are not unreasonable.

TABLE 2

CHANGE IN FCC VIEWS DUE TO CRITICAL ANALYSES BY INTERESTED PARTIES

Alternative	Prior Probability	$P(X \mid \text{Alternative})$	Posterior Probability
Helps	.32	.90	.31
Neutral	.43	.95	.43
Hurts	.25	1.00	.26

high. After, the probability distribution is much flatter. The probability of no significant effect is highest, but none of the alternatives has a probability of over .5.

I find this increase in uncertainty (both in the model and in the quotations) quite interesting. One ordinarily thinks of analysis as something that decreases uncertainty, something that sharpens our ability to make predictions. The model shows why this is not necessarily the case. When the conclusions of analysis conflict with strongly held prior views, the result may be to undermine confidence in the prior views without replacing them. The appearance of an analysis (even one with strong conclusions) may leave a decision maker (even a perfectly rational one) with a fuzzier view of the future than the one he started with.

The posterior probabilities from the table above become the priors for the next stage: consideration of the critical analyses submitted by the various interested parties, as shown in table 2. Here, the new information, X, is the appearance of the interested parties' criticisms of the staff study and the RAND reports. Everyone involved seems convinced that such criticisms would appear no matter what. The probabilities of X conditional on the alternatives are all high, and only slightly greater if cable really hurts UHF (1.00) than if cable helps UHF (.9).

Applying Bayes's theorem to calculate posterior probabilities, we see that the analyses submitted by the interested parties were too (unconditionally) predictable to have much effect on the FCC's views; the posterior probabilities in table 2 are almost the same as the priors. This does not mean that the interested parties' analytical efforts were wasted, however. For if the appearance of the studies was predictable, their nonappearance would have been a surprise with considerable information content. In terms of the model, if X is the nonappearance of critical studies by the interested parties, the conditional probabilities of X are complements of those shown in the table: $P(X \mid \text{helps}) = .1$, $P(X \mid \text{neutral}) = .05$, $P(X \mid \text{hurts}) = 0$. Bayes's theorem then yields posterior probabilities that are largest for helps (.6), while neutral is second most probable (.4), and hurts is out of the running (.0). If the interested parties had not submitted their critical analyses, the change in FCC views would have been much sharper than it actually was.

But the critical analyses were submitted, so the final posterior probabilities are those shown in the last column of table 2. These probabilities, calculated in the model, correspond very well to the verbal statement of FCC views in the postanalysis quotation above. The increased uncertainty about the exact outcome, the sharp decrease in probability that UHF would be significantly hurt, and the slight edge given to the probability of no significant impact are all evident in the quotation as well as in the model.

Conclusion

To avoid possible misunderstanding, I list a number of things that I do not do in this note:

1. I do not argue that the FCC actually made calculations using Bayes's theorem. Of course it did not, any more than a driver deciding whether it is safe to pass solves equations relating distances, speeds, and accelerations. But in both cases, the equations may help in understanding the subjective processes that actually take place.
2. I do not prove that policy analysis had any effect on FCC views. Analysis affects views in the model, and the model is consistent with available evidence. But presumably other models could be const explain the change in views without reference to policy analysis
3. I do not test the Bayesian model as a general representati effect of analysis on decision-makers' views. To do so would (among other things) observations on considerably more than case treated here.

On the other hand, I have presented (what seems to me to interesting framework for interpreting changes in decision-makers' views one that encompasses and in some sense explains several interesting aspects of the FCC's cable television proceedings. For example, it shows how policy analysis can increase uncertainty about the future, and it explains the apparently slight impact of studies commissioned by interested parties. In addition, it shows how decision-makers' views—quite rationally—can be influenced by analyses that they neither read nor understand.

References

American Broadcasting Company. "Comments in FCC Docket No. 18397-A." New York: American Broadcasting Co., December 7, 1970.

———, "Reply Comments in FCC Docket No. 18397-A." New York: American Broadcasting Co., February 10, 1971.

Association of Maximum Service Telecasters, Inc. "Comments in FCC Docket No. 18397-A." Washington: Assoc. Maximum Service Telecasters, Inc., December 7, 1970.

———. "Reply Comments in FCC Docket No. 18397-A." Washington: Assoc. Maximum Service Telecasters, Inc., February 10, 1971.

Charles River Associates Incorporated. "The Impact of CATV on Local Television Stations: A Critique of the FCC's Staff Report." Cambridge, Mass.: Charles River Assoc. Inc., September 1970 (Appendix C to Twenty-one Television Stations [1970]).

Dimling, John A., Jr. "Analysis of 'The Economics of the TV-CATV Interface' as It Relates to the Impact of Distant Signal Importation on Local Stations." Washington: Nat. Assoc. Broadcasters, December 7, 1970 (Appendix D to National Association of Broadcasters [1970]).

Federal Communications Commission. *Notice of Proposed Rule Making and Notice of Inquiry in Docket No. 18397.* FCC 68-1176, 15 FCC 2d 417, 33 Fed. Reg. 19028, 1968.

———. "The Economics of the TV-CATV Interface." Washington: FCC, Broadcast Bureau, Research Branch, July 15, 1970. (*a*)

———. *Second Further Notice of Proposed Rule Making in Docket No. 18397-A.* FCC 70-676, 24 FCC 2d 580, 35 Fed. Reg. 11045, 1970. (*b*)

———. *Letter of Intent: Cable Television Proposals.* 31 FCC 2d 115, August 5, 1971.

———. *Cable Television Report and Order in Dockets 18397, 18397-A, 18373, 18416, 18892, and 18894.* 37 Fed. Reg. 3252–3341, February 12, 1972.

Fischman, Leonard L. "Critique of Study by Rolla Edward Park on Potential Impact of Cable Growth on Television Broadcasting." Santa Monica, Calif.: Econ. Res. Assoc., February 1971 (Appendix A to Twenty-one Television Stations [1971]).

Kaiser Broadcasting Corporation. "Comments in FCC Docket No. 18397-A." Oakland, Calif.: Kaiser Broadcasting Corp., December 4, 1970.

M. H. Seiden and Associates, Inc. "CATV Distraction Effect." Washington: M. H. Seiden & Assoc., November 1970 (Appendix A to Twenty-one Television Stations [1970]).

National Association of Broadcasters. "Comments in FCC Docket No. 18397-A." Washington: Nat. Assoc. Broadcasters, December 7, 1970 (includes Dimling [1970] and Statistical Research Inc. [1970]).

———. "Reply Comments in FCC Docket No. 18397-A." Washington: Nat. Assoc. Broadcasters, February 10, 1971.

Park, R. E. *Potential Impact of Cable Growth on Television Broadcasting.* R-587-FF. Santa Monica, Calif.: RAND Corp., October 1970 (summarized in "The Growth of Cable TV and Its Probable Impact on Over-the-Air Broadcasting," *A.E.R.* 61, no. 2 [May 1971]: 69-73.)

———. *Cable Television and UHF Broadcasting.* R-689-MF. Santa Monica, Calif.: RAND Corp., January 1971 (also appears as "Cable Television, UHF Broadcasting, and FCC Regulatory Policy," *J. Law and Econ.* 15 [April 1972]: 207-31).

Statistical Research, Inc. "The Potential Impact of CATV on Television Stations." Westfield, N.J.: Statis. Res. Inc., Fall 1970 (Appendix E to National Association of Broadcasters [1970]).

Twenty-one Television Stations. "Comments in FCC Docket No. 18397-A." Washington: Covington & Burling, December 7, 1970 (includes Charles River Associations [1970] and M. H. Seiden and Associates [1970]).

———. "Reply Comments in FCC Docket No. 18397-A." Washington: Covington & Burling, February 10, 1971 (includes Fischman [1971]).

On Some Unhappy Consequences of Professor Liebhafsky's Consumers' Demand for Psychiatric Help

Jean-Claude Koeune

University of Minnesota

In a recent article in this *Journal*, Liebhafsky (1972) exhibits a utility function that generates, for one of the two goods, a perfectly price-inelastic demand schedule. Specifically, the utility function is

$$U = \log_e x_1 + \frac{p_2^2}{Ip_1} x_2,$$

(I is money income), and the resulting demand functions are

$$x_1 = \frac{I}{p_2} \quad \text{and} \quad x_2 = \frac{I(p_2 - p_1)}{p_2^2}.$$

Thus, the demand for X_1 will not depend upon its own price, tho
is a normal good with unitary income elasticity.

In addition, in view of the rather unusual features imposed on the utility function in order to reach the desired result, Liebhafsky undertakes to provide "a realistic explanation" for it: "The consumer," we are told, "suffers from a guilt complex arising out of the fact that he is one of the affluent members of society" (p. 831). This explains why prices and money income are present in his utility function: "Changes in these independent variables have a psychological effect on the consumer in addition to the effect they exert indirectly through the goods consumed" (ibid.). To alleviate his guilt complex, the consumer buys psychiatric help, represented by X_1. An ad hoc explanation of the fact that his demand for psychiatric help will be price inelastic is also provided, as well as some policy conclusions, but this is accessory.

The purpose of this comment is twofold: first, to argue that some features of the utility function, while crucial to the attainment of the desired price-inelastic demand function, are inconsistent with the

[*Journal of Political Economy*, 1974, vol. 82, no. 5]
© 1974 by The University of Chicago. All rights reserved.

"realistic explanation" that underlies it; and second, to show that even if one accepted the utility function as it stands, the demand for psychiatric help would not be price inelastic in all circumstances, and would not remain price inelastic for very long.

The negative direct influence of money income on utility is explained by the fact that, being an affluent member of society, the consumer suffers from a guilt complex. One may reasonably ask why. If he is concerned by his less affluent fellowmen's welfare, this concern ought to appear in some way in his utility function. Of course, one could retort, this is indeed the underlying rationale, and it appears in the utility function precisely through the introduction of money income, with a negative effect, among the arguments: this allows doing away with the more cumbersome notation that explicit consideration of interdependent utility functions would inevitably entail. However, if the reason for the affluent consumer's feeling guilty is some interdependence among utilities, one should also, to be consistent, consider that there are other ways to alleviate the bad feelings: income redistribution through charity or other means comes immediately to mind (or even trading some income for more leisure). But this means that there exist substitutes for psychiatric help, which makes it harder to justify that x_1 will not respond to a change in its price.

Still, let us accept the presence of money income in the utility function. What about p_2? "A decrease in the price of X_2," it is said, "has the same effect in increasing the consumer's feeling of guilt as an increase in income" (p. 831). This is difficult to admit because, other things being equal, a decrease in p_2 should increase real income relatively more for the less affluent consumers than for the more affluent whose utility function is under scrutiny: after all, while the assumption is not explicitly made, we are certainly led to believe that less affluent consumers will not need psychiatric help because they will not have any guilt complexes to cure: therefore, consuming X_2 only, they will benefit relatively more from a reduction in p_2. Why this should increase the affluent consumer's feeling of guilt is a mystery.

There is a symmetrical argument to be made with respect to the presence of p_1, with a negative influence, in the utility function: other things being equal, an increase in p_1 will reduce real income more for the consumer of X_1 than for nonconsumers. That this should increase his guilt feelings appears inconsistent once again. On the other hand, as can readily be seen, if p_1 is dropped from the utility function, the demand function for X_1 will no longer be price inelastic.

Suppose, however, that, all reservations being made, one accepts the utility function contrived by Liebhafsky. Does it follow that the demand for X_1 will always be price inelastic? The answer is negative. Consider the demand function for X_2:

$$x_2 = \frac{I(p_2 - p_1)}{p_2^2}.$$

"With $p_2 > p_1$," says the author, "X_2 is a normal commodity and has a negatively sloping demand curve" (p. 831). First, let us note in passing that the second part of the statement does not follow from the proviso: one can easily verify that the elasticity of x_2 with respect to its own price is

$$\frac{2p_1 - p_2}{p_2 - p_1}.$$

Thus, for $2p_1 > p_2 > p_1$, X_2 will be a normal commodity with a positively sloping demand curve. But what happens for $p_2 \leq p_1$? Liebhafsky sweeps this unpleasant prospect away by claiming that "we can always, in principle, choose a unit of measurement of x_1 sufficiently small (approaching zero as a limit) to guarantee that $p_1 < p_2$" (p. 830). Unfortunately, this will not work! The reason is quite simple: no matter what units we initially select for measuring X_1, X_2, p_1, and p_2 in such a way that p_2 is greater than p_1 by a comfortable margin, we cannot foreclose the possibility for a change in prices to bring p_1 above p_2 (especially assumption that "the psychiatrist has managed to produce differentiated product". What shall we do then? Can we simp the unit of measurement of X_1 to insure that its unit price becom smaller than p_2? But we cannot do this without at the same time the writing of the utility and demand functions,[1] unless we ar admit that the consumer's utility and demands for goods depend , which units are chosen to measure the quantities of goods: then it is obvious that he needs psychiatric help! Thus, once p_2 becomes smaller than p_1, there is no way to avoid or postpone that annoyance.

---

[1] Specifically assume that $y_1 = Ax_1$, where y_1 represents the quantity of good X_1 measured in some other units, and A is a positive number (e.g., $A = 60$ if y_1 is the quantity of psychiatric help measured in minutes of the psychiatrist's time, and x_1 represents the same quantity measured in hours). If q_1 is the price of the new unit of X_1, one necessarily has

$$q_1 = \frac{p_1}{A}.$$

The utility function will now be written

$$U = \log_e\left(\frac{y_1}{A}\right) + \frac{p_2^2 x_2}{AIq_1}$$

and the demand functions will now be

$$y_1 = \frac{AI}{p_2} \quad \text{and} \quad x_2 = \frac{I(p_2 - Aq_1)}{p_2^2}.$$

What happens then? Since x_2 cannot be negative, one gets a corner solution, that is

$$x_2 = 0, \qquad \text{and} \qquad x_1 = \frac{I}{p_1}$$

and the demand function for X_1 becomes price elastic.[2]

What is the likelihood of p_1's becoming greater than p_2? Very great indeed, if the psychiatrists are profit maximizers (or effort minimizers). The mind boggles at the thought of what could happen! Notice that the total revenue the psychiatrist extracts from his affluent consumer is $p_1 x_1 = I p_1 / p_2$ as long as $p_1 \leq p_2$, and simply I for $p_1 \geq p_2$, since from there on the consumer will only consume psychiatric help. Thus, the psychiatrist maximizes his income (capturing all of his customer's income) by setting p_1 at least equal to p_2, and he minimizes effort by increasing p_1—and thus decreasing x_1—up to the point where he gives his affluent customer just enough psychiatric help to deter him from committing suicide. The consumer will starve to death anyway, since he consumes nothing but psychiatric help, but never mind. Once the most affluent members of society have died in their psychiatrist's arms, the slightly less affluent become the most affluent, they start having guilt complexes, and rush to the nearest shrink's office. In the long run, they are all dead, save the psychiatrists, who may start feeling guilty about something.

Reference

Liebhafsky, H. H. "The Rational Consumer's Demand for Psychiatric Help: A Preference Function Generating a Perfectly Price-inelastic Demand Function." *J.P.E.* 80 (July/August 1972): 829–32.

[2] One can easily verify that the corner solution is truly a maximum: the expression for the marginal rate of substitution (*MRS*) given by the utility function is

$$\frac{U_1}{U_2} = \frac{I p_1}{x_1 p_2^2}.$$

At the corner, $x_1 = I/p_1$, and the value of *MRS* becomes simply $(p_1/p_2)^2$. This is greater than the price ratio if $p_1 > p_2$, but one cannot bring about the equality by increasing x_1 and decreasing x_2, since x_2 is zero. On the other hand, at the corner, *MRS* is smaller than the price ratio if $p_1 < p_2$. One can bring about the equality and increase total utility by decreasing x_1 and increasing x_2 from zero to some positive value.

Book Reviews

Discrimination in Labor Markets. Edited by Orley Ashenfelter and Albert Rees. Princeton, N.J.: Princeton University Press, 1973. Pp. xi+181. $9.00.

This volume contains five papers originally presented at a conference on discrimination in labor markets, held at Princeton University in October of 1971, along with the formal comments of the conference discussants. It starts with a theoretical piece by Kenneth Arrow: continues with three empirical studies by Finis Welch, Orley Ashenfelter, and Ronald Oaxaca; and concludes with a paper on government policy by Phyllis Wallace. The papers are all nontechnical and designed to appeal to a wide audience of readers. While more technical versions of the first four papers have recently been published in professional journals or other books, the editors have performed a valuable service by collecting the expository versions in one place.

Kenneth Arrow's paper discusses the theoretical foundations of the neoclassical analysis of labor-market discrimination. Arrow elaborates on the pioneering work of Gary Becker and explicitly includes general equilibrium consideration
Several models with employers or employees having "tastes for discrim
are developed, and the conditions under which wage differentials and
gation will occur between groups of equally productive individuals of
races (or sexes) are stated. As pointed out by Melvin Reder in his com
the volume, these models all assume completely wage-inelastic lal
curves, and hence they neglect the possibility of labor-market discrimin
influencing the aggregate group-employment levels.

Arrow concludes his paper by extending the models to consider costs of
ment and statistical discrimination as possible explanations for why labor-mark
discrimination persists over time. The later approach depends crucially on the existence of costs to employers of acquiring information on individual applicants' productivity and on employers perceiving that on average different race (or sex) groups vary in their productivity. Arrow shows quite strikingly that even if this perception is incorrect at a point in time, statistical discrimination may lead the "discriminated-against" group in fact to invest relatively less than the other group in increasing their productivities in future periods. This point has been forcefully made also by Michael Spence (1973) in his recent work on job-market signaling.

The next three studies summarize econometric studies of various facets of labor-market discrimination, with each based upon data on individuals taken from the 1967 *Survey of Economic Opportunity* (*SEO*) and/or the one in 1,000 sample of the 1960 census. Each paper contributes significantly to our knowledge of the subject.

The relationship between education and racial discrimination is considered by Finis Welch in his contribution. Previous econometric studies, based upon cross-section data for a single year, have uniformly found lower rates of return to education for blacks than whites. However, Welch suggests that these studies may be misspecified because they group different cohorts of individuals together and neglect the substantial increases in the relative "quality" of blacks' education (as compared to whites' education) that have occurred historically. Indeed, using data from both the 1967 *SEO* and the 1960 census together, his estimates indicate that rates of return to education for blacks have risen over time relative to the comparable rates of return for whites. Moreover, for recent cohorts of students, the former may actually exceed the latter.

Orley Ashenfelter seeks in his paper to measure the impact of trade unionism in the American economy on black/white and female/male income differentials. He argues a priori, and his empirical results confirm that industrial unions have raised and craft unions lowered the average wage of black workers relative to that of white workers, as compared to the analogous differential in nonunion labor markets. However, Ashenfelter finds that for the economy as a whole, the impact of unions on income differentials by race and sex is extremely small. He calculates that unions have reduced the overall black/white income differential and the comparable white female/white male income differential by less than 2 percent. That is, trade unions do not appear substantially to have exacerbated the extent of labor-market discrimination that would exist in the absence of unions.

In the final econometric study in the volume, Ronald Oaxaca considers sex discrimination in the labor market. Oaxaca specifies and estimates wage equations for males and females based upon the postschooling investment model of human-capital formation. He shows that even after variables that reflect individuals' productivity and measures of job experience are controlled for, a large component of the observed female/male wage differential remains unexplained and may be attributable to labor-market discrimination. Oaxaca emphasizes, as does Barbara Bergmann in her conference comments and previous research, that occupational barriers that women face are a major source of the observed differential. Furthermore, these same barriers may prevent females from investing as much in on-the-job training as do males.

The volume concludes with a useful paper by Phyllis Wallace that examines public policy toward discrimination by race and sex. Wallace discusses how legal mechanisms have been quantified into operating policies, and, relying upon the limited experience which we have with these policies, suggests alternative strategies. Unfortunately, as Wallace realizes, it is difficult to assess the effectiveness of government policies in the area solely by looking at input variables, such as the number of complaints processed by the Equal Employment Opportunity Commission, as she does. A more analytical approach would be to attempt to relate statistically, *ceteris paribus*, the magnitudes of the policy input variables to desired outcome variables (such as wage rates, employment levels, or occupational mobility of nonwhites and females) which the policies are designed to influence. Recent works by Orley Ashenfelter and James Heckman (forthcoming) using individual establishment data and Richard Freeman (cited in his comments in this volume) using aggregate data are examples of this type of approach and add considerably to our knowledge of the effectiveness of these policies.

In sum, I consider *Discrimination in Labor Markets* a fine volume. Anyone who has the slightest interest in the analysis of labor-market discrimination should seriously contemplate purchasing it. The relatively nontechnical nature of the papers will appeal to a wide range of readers, and the book should quickly find its way onto reading lists for undergraduate and graduate courses that discuss the economics of discrimination.

Ronald G. Ehrenberg

University of Massachusetts, Amherst

References

Ashenfelter, Orley, and Heckman, James. "Measuring the Effect of an Anti-Discrimination Program." Paper presented at a conference on "Evaluating the Labor Market Effects of Social Programs," Princeton, N.J., May 2, 1974.

Freeman, Richard. *Black Elite: Education and Labor Market Discrimination.* New York: McGraw-Hill, 1973.

Spence, Michael. "Job Market Signaling." *Q.J.E.* 87, no. 3 (August 1973): 355–74.

The Economy of Love and Fear: A Preface to Grants Economics. By KENNETH E. BOULDING.
Belmont, Calif.: Wadsworth Publishing Co., 1973. Pp. v + 116.

When I first heard an exposition of the theory of grants economics by Martin Pfaff, in the presence of Professor Boulding, I was injudicious enough to express the opinion that "grants economics" was merely a new terminology for some already recognized problems in economics, and that it had yet to prove its usefulness by producing propositions and insights that were novel and convincing. I undertook to review this book, the first of a projected series, in order to give grants economics a second chance. Unfortunately, it fails equally miserably the second time around. The major reason, bluntly, is that Boulding, having left pure economics in 1943 to become a labor economist, thence to become interested in the social system, conflict and conflict resolution, and the "integrative sy leading into grants, has returned to economics more or less as he left it 30 y

More accurately, a caricature of economics as it was 30 years ago support the general claim advanced in the *Introduction to the Series* th economics is the answer to criticisms of exchange economics by radical e and to the analytical weakness of radical economics itself, and to bu detailed claims of the text that economics has ignored or neglected a problems that it has not in fact neglected. For the same reason, the suggestions offered to demonstrate the insights of grants economics are remarka for their naiveté, ignorance, or sheer effrontery.

Thus, it is not very persuasive to be told that the family is a grants economy, so that in a pure exchange economy the family and society itself would disappear. Nor does it illuminate to be told, with respect to inflation, that prices and wages, and especially price dynamics, set up strains—and to have Gardiner Means's proposal for a tax on price and wage increases recommended for serious consideration. And it is surprising to be told that economists have more or less simply ignored the income redistributions consequent on any kind of change. We are left with the message that grants do too exist and the impression that any idea that recognition of their existence suggests to Boulding has necessarily escaped the attention of the rest of us.

There are two mildly interesting chapters on the concept of exploitation and on the limits of welfare economics as a guide to the problems of maintaining a functioning society, sandwiching one expressing Boulding's well-known views on armaments and deterrence. The final, cosmic chapter gives the word on spaceship *Earth* and expresses the hope that grants economics may lay the foundations for a new ideology of spaceship living in loving harmony to replace previous ideologies based on unlimited expansion.

Probably there are some first-year students trained in the more arid textbooks and a few young instructors from the more mathematical graduate schools, who will benefit from Boulding's reminder that society exists, and a larger number of self-styled radical economists patient enough to think about his message that grants have limits of acceptability on both sides. There are probably also a number of amateur-philosopher undergraduate teachers who will enjoy playing with his concepts for extending grants theory to parallel exchange theory, and some who will emphasize his errors in treating the GNP and its distribution and in outlining macroeconomics theory. But professional economists who know their fields will derive little nourishment from this book.

HARRY G. JOHNSON

University of Chicago

The China Trade and U.S. Tariffs. By HARRY A. CAHILL.
New York: Praeger Publishers, 1973. Pp. vi + 164. $15.00.

One of the most important questions posed in discussions concerning the extension of most-favored-nation (MFN) tariff rights to the People's Republic of China is the possible impact this would have on domestic producers. Would not the granting of the MFN "low" tariff to Chinese goods expose the U.S. market to a flood of those goods, inasmuch as China is a labor-surplus (low-wage) economy with a state monopoly not only over the allocation of resources to the production of export commodities, but their distribution among alternative trading partners as well? While a reduction in tariffs would increase the gains to be made, the policy option of dumping was made available to the Chinese when the United States unilaterally removed the legal obstacles to direct trade between the two countries in the summer of 1971. The limited evidence available thus far certainly shows no indication of their choosing such a policy, and their own experience in trade with the non-Communist countries, as well as the experience of others in trade with the United States, should make clear to them that even the slightest suspicion of dumping practices merely leads to retaliation.

The central argument of Harry A. Cahill's new book, *The China Trade and U.S. Tariffs*, gives unintended succor to those who argue against the extension of MFN status to the People's Republic of China. This support is unintended because the author's purpose is to counterbalance the arguments of the so-called pessimists in another debate concerning Sino-American trade: the debate over the potential volume of that trade. Admitting the direct trade which has occurred in the initial period following the opening of trade between these two countries was not large and, therefore, supports the arguments of the pessimists, Cahill attempts to show in this book that "this flow will remain constricted [only] as long as tariff treatment equitable with that accorded [China's] competitors is not given" (p. 69).

A brief summary of the history and recent changes in U.S. controls on trade with China (chap. 2) and the changes in the level and structure of China's economy (chap. 3) and foreign trade (chap. 4) since 1949 are provided as background information. Although the estimates published in U.S. government sources (12 of the 13 tables presented in chaps. 3 and 4 are from three such sources, which are readily available) are uncritically accepted, the summary presented in these chapters does not conflict with that well known by those readers who will be most interested in the much more specialized discussions in the remaining parts of the book.

Three alternative types of estimates which have been made for the potential trade between China and the United States are evaluated and a fourth possible approach suggested in chapter 5. The historical approach which argues U.S. trade with China in the future will be relatively small because it was a relatively small share of China's trade in the past (15–19 percent) is ruled out because the commodity composition of the two countries' trade has changed significantly between the pre- and the post-1949 periods. My own estimates, said to be based on arithmetic reasoning and the scientific method, are not ruled out but are questioned because of the need to rely on several crucial assumptions that may not be realistic. Businessmen, who became pessimistic after the disappointing failure of their optimistic hopes in the initial period of possible trade with China, are now said to make only cautious estimates or none at all. Cahill proposes a fourth approach—a pragmatic approach.

In the belief that the 15 months following the opening of trade between China and the United States provides us with a sufficient sample, Cahill presents a series of interesting accounts of U.S. businessmen's experiences, their opinions, and his own intuitive judgments (chap. 6) and reaches the following conclusion: this informal sample, supported by a sample of import declaration forms (nine purchasing orders with an average value of $7,500), indicates the widespread interest in the United States in a wide range of Chinese goods, but the failure of Chinese exports to the United States to grow more rapidly is mainly due to poor merchandising practices on the part of U.S. businessmen and the existence of the relatively high non-most-favored-nation (NMFN) tariff. While I do not agree with his interpretation of the evidence presented, nor the logic used to reach his conclusion, it can be accepted as a testable hypothesis.

To test the hypothesis, Cahill identifies almost 200 Chinese commod
believes have a sales potential in the U.S. market and presents a series of
statistical tables showing the total U.S. import of those commodities
sources and from the leading Asian sources in 1971 and the disparity betw
MFN and NMFN tariffs in each case. While no quantitative estimate
these statistics and the explicit arguments found throughout the book are
to counterbalance the estimates of us "seasoned analysts" who are p
about the potential growth of Sino-American trade.

There are several questions one could raise against Cahill's analysis, such (*a*) although mentioned, no serious consideration is given to the constraints on China's supply of these export commodities, nor is any evidence given to show the Chinese actually export these commodities to anyone; (*b*) although U.S. demand for these commodities is said to exist and the United States does import these commodities, if we assume Canadian tastes are similar to those in the States, China's exports to Canada in the absence of a discriminatory tariff have remained very small, even though Canada has a very large export surplus in trade with China; and (*c*) the total amount of U.S. imports of these approximately 200 commodities from non-Japanese, Asian sources in 1971 was less than U.S. $500 million and, if we make a very liberal assumption that China could obtain one third the market share of these sources when granted MFN tariff rights, Chinese exports of these commodities to the United States would amount to only U.S. $150 million which is well within my and other pessimistic estimates for potential Sino-American trade. Thus, not only do I disagree with the methodology and arguments Cahill uses to support his conclusion concerning the large volume of potential exports from China to the United States, I do not believe that conclusion follows from the methodology and arguments he presents.

My own disagreement with Cahill's methodology and arguments is of little importance, however, for his whole presentation can be questioned in light of the momentous developments which have occurred in Sino-American trade in the past year. While Cahill chose to concentrate on an analysis of potential Chinese exports to the United States, attention has shifted to the Chinese import side due to the decision of the Chinese to make very large purchases of agricultural products in the States. Sino-American trade will probably increase from U.S. $110 million (well within the pessimistic estimates) in 1972 to $800–$900 million in 1973 (well beyond the expectations of the pessimists). This level of trade probably will leave the United States as China's second leading trading partner with contracts for U.S. exports to China of grain, cotton, soybeans, and vegetable oils alone accounting for over two-thirds of the total trade turnover.

Cahill cannot be faulted for not anticipating this development, as no one else did either. The attempt to present his arguments to the public as quickly as possible when there was considerable interest in this topic, even though rapid changes were still occurring in the political and economic relations between the United States and China, however, had two unintended consequences. The first was an unavoidable one. The book was poorly edited in an attempt to have it published quickly (e.g., the state was said to control foreign exchange holidays instead of foreign exchange holdings, and the share of processed goods in total exports was said to increase from 28.9 percent in 1955 to 25.7 percent instead of 38.7 percent in 1965), but also quickly outdated by the rapid changes taking place (e.g., the only appendix is a detailed list of commodities which can be exported to the People's Republic of China under a general license, but this list was issued on August 10, 1971 and was replaced by a new list in February 1972, which had the effect of putting exports to China on an equal basis with exports to the Soviet Union and most other East European countries). The second consequence is much more important. I strongly agree with those who call for the extension of MFN rights to the People's Republic of China, but Cahill's analysis in support of that argument, which I have tried to show has some serious weaknesses, will only lend support to those who oppose the argument which both he and I support.

Robert F. Dernberger

University of Michigan

World Trade and Payments: An Introduction. By Richard E. Caves and Ronald W. Jones.
Boston: Little, Brown & Co., 1973.

As one would expect, Caves and Jones have combined to provide an excellent textbook. It should be seriously considered as a candidate for adoption in any undergraduate course in international economics, although its appeal will be greatest to those instructors who prefer a moderately heavy dose of theory.

The coverage and quality of basic theory is outstanding—it is hard to think of any major topics in trade and payments theory that are omitted. A priori, the danger of such extensive coverage is that emphasis may be lost, but Caves and Jones have managed to provide perspective. Moreover, the quality of the exposition is outstanding; many subjects apparently too difficult for an undergraduate course appear suitable with the Caves-Jones treatment.

The book is organized in a way that gives considerable flexibility to the instructor in terms of the range and depth of topics covered. The presentation in the text is geometric and verbal, but there are supplements for chapters covering theoretical material to provide a mathematic treatment of the same issues. The book is divided into six parts, the first four covering the main material. The first part covers the pure exchange model, the second gives trade and production. These two parts contain most of the subjects usually dealt with in "pure trade" theory. By the end of the second part, the student has been exposed to stability conditions, the factor price equalization, Stolper-Samuelson, and Rybczynski theorems, the Leontief paradox, technical change, and of course, the Ricardian model. Part 3 deals with commercial policy, including the optimal tariff, other arguments for and effects of tariffs, including optimal intervention in the presence of distortions, and customs union. The fourth part is entitled "Macroeconomic Concerns: Internal and External Balance," and covers balance-of-payments theory, international liquidity, and related issues of the international monetary system and adjustment. Parts 5 and 6 are on international factor movements (the transfer problem, portfolio and direct investment, and short-term capital flows) and trade-and-growth issues; these subjects are, as noted above, treated less extensively than the others and are not integral to the book. If there is a substantive omission of importance, it is probably underemphasis on trade-and-development issues.

Within each of the major parts there is attention to the underlying theory and to empirical testing of alternative models. There is deliberately no systematic account of the history of the international economy, on the ground that such an exercise "offers windy speculation rather than useful and empirically tested models" (p. 4). The objective is generally well carried out, although there are times when one feels that the authors have had to strain to label things empirical tests, as when they discuss testing the Heckshher-Ohlin theorem against historical patterns and conclude that "the Heckshher-Ohlin model explains historical patterns of trade and their changes" (p. 204), which is surely a bit strong.

Relative to positive theory and empirical testing, the Caves-Jones coverage of normative trade theory is somewhat sparse and not quite as For example, in the chapter on tariff theory, only two pages are giv efficiency of free trade," and focus there is on the contrast between wo and individual country welfare; exchange-control systems are not disc in the context of the quota-equivalent of tariffs and nontariff barriers In contrast to other textbooks, the change of emphasis from normative is welcome, although it may have been slightly overdone.

Moreover, there is occasionally a certain disquieting assumption of econom rationality on the part of governments. Discussion of stability conditions in the pure exchange model comes early, in chapter 3. After being introduced to the Marshall-Lerner condition in a barter model, the student is told that "the market in internationally traded commodities can have unstable equilibria. . . . Countries may attempt to *peg* their exchange rates because they fear an underlying instability in the market for foreign exchange that would lead to wide swings in currency values as a consequence of any small jiggle in the market" (p. 42). While such an outcome is of course possible, it is not clear that the possibility empirically deserves the prominence given to it, nor, for that matter, that rational concern about the likelihood of market instability is the motivating factor dictating pegged exchange rates. The discussion about international reserves is similarly marred, as the entire treatment is based on the implicit assumption that governments' holdings of reserves are based upon careful calculation: "Countries likewise hold reserves to see them through alternating periods of surplus and deficit" (p. 415).

These objections are of second-order importance, however. Caves and Jones have done the profession a great service by providing a first-rate textbook on international economics. By appropriate selection of material, it is suitable for use at any undergraduate level from first course after principles to honors seminar.

ANNE O. KRUEGER

Massachusetts Institute of Technology and University of Minnesota

In Pursuit of Price Stability: The Wage-Price Freeze of 1971. By ARNOLD R. WEBER. Washington: Brookings Institute, 1973.

The wage-price freeze announced by President Nixon on August 15, 1971, was probably the most important shift in U.S. economic policy since the end of World War II. But as of this writing—January 1974—it is clear that the economic stabilization program has not proved to be stabilizing. The question is why. The answer may have much to do with whether a future U.S. government again resorts to direct controls.

Those personally involved in designing and administering the stabilization program can contribute enormously to our understanding if they will follow Arnold Weber's lead in reporting what happened and their views as to why the stabilization program took the form it did. Weber, as executive director of the Cost of Living Council during Phase 1, was in a position to know most of the inside story. No economist or political scientist with a serious interest in the economics and politics of controls can afford not to read this well-written and informative book.

The first two chapters of Weber's book contain a discussion of the political and economic background leading up to the freeze and some relatively brief general comments on incomes policies. A description of the events and problems of the freeze takes up the bulk of the book, chapters 3–8. Much of this material will be familiar to those who read the newspapers faithfully throughout the freeze. Nevertheless, Weber's discussion is enlightening because he resists the temptation simply to record events chronologically and instead reports events that were significant in shaping the program or that illustrate general features of the freeze. In the final chapter, "Lessons and Limitations," Weber presents his overall views on the experience.

In this final chapter Weber does not make the generalization that seems most obvious to this reviewer: since the freeze could not be successful without the active cooperation of the general public, all of the individual policy decisions were conditioned by the need to retain public support. Whenever there was conflict between economic efficiency or individual equity and the maintenance of public support, policy was determined by the latter rather than the former. Law and economics were highly relevant, but more as inputs to predicting the secondary repercussions of, and public reactions to, policy decisions than as guiding principles of policy.

The Florida Telephone dividend case provides a clear example of the overriding importance of the public relations aspects of freeze policymaking. Although there was no legal authority to control dividends, there was a voluntary appeal for all companies to hold the line on their payouts.

> Only one company, the Florida Telephone Corporation, resisted the government's blandishments. Both the president and attorney of the firm rejected the implication that their corporation had committed any illegality or impropriety, and they attacked the summons to Washington as an improper exercise of power. With each of the other companies remedies were worked out which brought actual or symbolic compliance with the freeze on dividends.
> Only in the case of the Florida Telephone Corporation was additional leverage necessary. Hence, contacts were made by administration officials with members of the financial community and large customers of the company who presumably could exert some pressure on management. [Pp. 93–94]

The government won.

Weber is obviously troubled by the Florida Telephone and other such cases.

> The sense of bureaucratic achievement should not obscure a serious deficiency in the conduct of the freeze. Private economic decisions are essentially the exercise of individual liberty supported by law. Normal government interventions into the product and labor markets are conditioned by the requirements of due process arduously developed over many years. Because of time constraints, the sweeping application of the wage-price freeze, and the paucity of procedural guides or standards in the Economic Stabilization Act, important elements of due process were undermined in the course of administering the program. During the freeze, prior publication of rules in the *Federal Register*, hearings and appeals procedure, and the other niceties of due process generally were ignored. This helped to create a strong sense of grievance among several groups—particularly organized labor—and made it difficult to enlist the continuing cooperation necessary for a long-term program. Some short-circuiting of due process is necessary and probably inevitable in the imposition of a wage-price freeze. With prior planning, however, at least rudimentary safeguards can and should be established to protect individual rights and to minimize the damage to existing institutional arrangements, such as collective bargaining and freedom of contract, that have been traditionally sanctioned by society. [P. 129]

This statement is unexceptional, but can the reader be forgiven for asking about the details of the "rudimentary safeguards [that] can and should be established"? Is the maintenance of due process really feasible in the circumstance of a sudden freeze? I would contend that there is no possibility of setting up an administrative machinery to sort out the equities of the millions of contracts that are rewritten or broken by a wage-price freeze.

In a number of places Weber emphasizes that the bureaucracy that and administered the freeze was very small. While his interest is in sho the job can be done, at least for a time, with a small bureaucracy, I wo that it really is quite amazing that a highly complex and decentralized manned by an individualistic people could be so controlled. The pos mobilizing public support in this fashion must be regarded as a great for a free society; without the certainty of obtaining quick voluntary cooperation to deal with emergency situations (such as natural disasters threats of war) requiring community action, it would probably be necessary to maintain on a permanent basis much more extensive government control than has been customary in the United States. Although people are clearly willing to give up due process and a lot more in an emergency, it is not possible to mobilize massive support for a new program every month. Was the wage-price freeze a wise use of this valuable public resource?

As a final comment, I would note that there must be a lot more to the inside story than Weber has disclosed. In chapter 1, "The Background," he provides a clear discussion of the disappointments in the economic picture before the freeze and makes frequent reference to the importance of political factors. There is, however, no mention of the fact that the November 1972 election was not all that far away. Moreover, while Weber explicitly outlines the prefreeze views of the chairman of the Federal Reserve Board, Arthur Burns, the views of the other personalities involved are those of anonymous "chief economic advisers." Weber does not discuss who argued for what, or even, leaving personalities aside, what options were debated before the decision to impose the freeze was made. Did not someone ask: "What happens after the freeze? Where will it all end?"

WILLIAM POOLE

Federal Reserve Bank of Boston

SCOTTISH JOURNAL OF POLITICAL ECONOMY

Published three times a year in February, June and November

The Journal is aimed particularly at economists working in the fields of APPLIED ECONOMICS and ECONOMIC POLICY. Articles on related economic theory are also featured. Some recent contributions indicating the range of topics covered include:

A Mixed Strategy for Labour Economists	***JR Crossley***
Measuring the Labour Market Costs of Housing Dislocation	***H Kasper***
Inflation in the Context of Relative Deprivation and Social Justice	***JL Baxter***
Foreign Trade and British Economic Growth	***RCO Matthews***
Technology Gaps between Rich and Poor Countries	***Paul Streeten***
Shift-Share Analysis as a Guide to the Employment Performance of West-Central Scotland	***JN Randall***
Management Behaviour and a Theory of Diversification	***CJ Sutton***

The Journal includes a REVIEW SECTION containing review articles by leading authorities in their fields which appraise the contribution of recent publications in selected areas. Among the areas recently reviewed are the following.

Regional Economics	***Harry W Richardson***
Incomes Policy	***G Routh***
Industrial Organization	***H Townsend***
Development Economics	***P Robson***
Government and Planning	***A Nove***

Subscriptions £5, USA $17.50 should be remitted to Longman Group Limited, Journals Division, 43/45 Annandale Street, Edinburgh, EH7 4AT, Scotland.

A Theory of Social Interactions

Gary S. Becker

University of Chicago and National Bureau of Economic Research

This essay uses simple tools of economic theory to analyze interactions between the behavior of some persons and different characteristics of other persons. Although these interactions are emphasized in the contemporary sociological and anthropological literature, and were considered the cornerstone of behavior by several prominent nineteenth-century economists, they have been largely ignored in the modern economic literature. The central concept of the analysis is "social income," the sum of a person's own income (his earnings, etc.) and the monetary value to him of the relevant characteristics of others, which I call his social environment. By using the concept of social income, I can analyze the effect on these expenditures of changes in different sources of income and in different prices, including the "price" of the social environment. Interactions among members of the same family receive the greatest attention. The "head" of a family is defined not by sex or age, but as that member, if there is one, who transfers general purchasing power to all other members because he cares about their welfare. A family with a head is a highly interdependent organization that has the following properties: A redistribution of income among members does not affect the consumption or welfare of any member because it simply induces offsetting changes in transfers from the head. Not only the head but other members too act "as if" they "loved" all members, even when they are really selfish, in the sense that they maximize not their own income alone but family income. Transfers from parents to children in the form, say, of schooling, gifts, and bequests tend to be negatively related to what the income of children would be relative to their parents in the absence of these transfers. Therefore, the relative income of children *inclusive* of transfers could be unrelated or even negatively related to these transfers. Consequently, one cannot infer anything about the stability across generations of economic or social positions simply from knowing the relation between parental position and the amount transferred.

[*Journal of Political Economy*, 1974, vol. 82, no. 6]

© 1974 by The University of Chicago. All rights reserved.

"No Man is an Island." [JOHN DONNE, *Devotions upon Emergent*]

"Man is a social animal." [SENECA, *De beneficiis*]

1. Introduction

Before the theory of consumer demand began to be formalized by Jevons, Walras, Marshall, Menger, and others, economists frequently discussed what they considered to be the basic determinants of wants. For example, Bentham (1789, chap. 5) discusses about 15 basic kinds of pleasures and pains—all other pleasures and pains are presumed to be combinations of the basic set—and Marshall (1962, bk. 3, chap. 2) briefly discusses a few basic determinants of wants before moving on to his well-known presentation of marginal utility theory. What is relevant and important for present purposes is the prominence given to the interactions among individuals.

Bentham mentions "the pleasures . . . of being on good terms with him or them," "the pleasures of a good name," "the pleasures resulting from the view of any pleasures supposed to be possessed by the beings who may be the objects of benevolence," and "the pleasures resulting from the view of any pain supposed to be suffered by the beings who may become the objects of malevolence." Nassau Senior said that "the desire for distinction . . . is a feeling which if we consider its universality, and its constancy, that it affects all men and at all times, that it comes with us from the cradle and never leaves us till we go into the grave, may be pronounced to be the most powerful of all human passions" (quoted by Marshall 1962, p. 87.) Marshall also stresses the desire for distinction and illustrates its influence by discussing food, clothing, housing, and productive activities.[1]

Over the years I have received helpful comments on a succession of drafts from numerous persons, especially my colleagues at the University of Chicago and the National Bureau of Economic Research. I received very useful comments on the draft prepared for publication from Robert Barro, Isaac Ehrlich, Sam Peltzman, and George Stigler, and valuable research assistance from Walter Wessels. My research has been supported by a grant to the NBER from the National Institute of Child Health and Human Development, National Institutes of Health, U.S. Department of HEW; but the paper is not an official NBER publication since it has not been reviewed by the NBER Board of Directors.

[1] He limits his discussion of consumer demand to the largely formal theory of marginal theory because of the importance he attaches to the interaction between activities, consumer behavior and the basic wants: "Such a discussion of demand as is possible at this stage of our work must be confined to an elementary analysis of an almost purely formal kind" (1962, p. 90). He never developed the more complicated and less formal analysis.

As greater rigor permeated the theory of consumer demand, variables like distinction, a good name, or benevolence were pushed further and further out of sight. Each individual or family generally is assumed to have a utility function that depends directly on the goods and services it consumes. This is not to say that interactions between individuals have been completely ignored. Pigou (1903), Fisher (1926, p. 102), and Panteleoni (1898)[2] included attributes of others in utility functions (but did nothing with them). In recent literature, "demonstration" and relative income" effects on savings and consumption,[3] "bandwagon" and "snob" influences on ordinary consumption theory,[4] and the economics of philanthropic contributions[5] have been discussed. But these efforts have not been unified and, more significantly, have not captured the dominance attributed to social interactions by nineteenth-century economists.

Of course, sociologists have for a long time emphasized the central role of interactions and their importance in the basic structure of wants or personality. Veblen's conspicuous consumption and conspicuous leisure (if for this purpose he is classified as a sociologist) have entered ordinary discourse. At one point he said: "But it is only when taken in a sense far removed from its naive meaning that the consumption of goods can be said to afford the incentive from which accumulation invariably proceeds The motive that lies at the root of ownership is emulation," an usual basis of self-respect is the respect accorded by one's ne (Veblen 1934, pp. 25, 30). Interactions were also empha Durkheim, Simmel, Freud, and Weber, as well as in modern di of "social exchange" and the "theory of action" (see Blau 1968 1968).

My interest in interactions can probably be traced to a study of dis ination and "prejudice" where I analyzed discriminatory behavior by incorporating the race, religion, sex, or other personal characteristics of employees, fellow workers, customers, dealers, neighbors, etc., into utility functions (Becker 1971 [1st ed., 1957]). Subsequently, in order to provide a theoretical framework for a study of philanthropy by the National Bureau of Economic Research, I incorporated the standard of living of "poorer" persons into the utility functions of "richer" ones (Becker 1961). Further reflection gradually convinced me that the emphasis of earlier economists deserved to be taken much more seriously because social interactions had significance far transcending the special cases discussed by myself[6] and others.

---

[2] I owe this reference to George Stigler.

[3] See, e.g., Brady and Friedman (1947), Duesenbery (1949), or Johnson (1952).

[4] See Leibenstein (1950).

[5] See Vickery (1962), Schwartz (1970), Alchian and Allen (1967, pp. 135–42), and Boulding (1973).

[6] Other drafts that were also circulated include Becker (1968).

This essay incorporates a general treatment of interactions into the modern theory of consumer demand. In Section 2, various characteristics of different persons are assumed to affect the utility functions of some persons, and the behavioral implications are systematically explored. Section 3 develops further implications and applications in the context of analyzing intrafamily relations, charitable behavior, merit goods and multiperson interactions, and envy and hatred. The variety and significance of these applications is persuasive testimony not only to the importance of social interactions but also the feasibility of incorporating them into a rigorous analysis.

2. Theoretical Framework

A. Equilibrium for a Single Person

According to the modern (and very old!) theory of household behavior,[7]

$$U_i = U_i(Z_1, \ldots, Z_m) \tag{2.1}$$

is the utility function of the ith person, and $Z_1, \ldots, Z_m$ are the basic wants or commodities. As indicated earlier, Bentham mentions about 15 basic wants, whereas Marshall and Senior stress an even smaller number. Each person also has a set of production functions that determine how much of these commodities can be produced with the market goods, time, and other resources available to him:

$$Z_j = f_j^i(x_j, t_j, E^i, R_j^1, \ldots, R_j^r), \tag{2.2}$$

where x_j are quantities of different market goods and services; t_j are quantities of his own time, E^i stands for his education, experience, and "environmental" variables; and $R_j^1, \ldots, R_j^r$ are characteristics of other persons that affect his output of commodities. For example, if Z_1 measures i's distinction in his occupation, $R_1^1, \ldots, R_1^r$ could be the opinions of i held by other persons in the same occupation. Presumably, characteristics of others affect the production of a significant fraction of commodities.

If the R_j were completely outside i's control—that is, unaffected by what he does with his resources—i would maximize U taking the R_j as given. This is one way to justify the usual neglect of interactions. They are considered beyond the control of the persons being studied and are therefore taken as given when one is analyzing their reactions to changes in resources and prices.

The point of departure of my approach is to assume the contrary, namely, that i can change R_j by his own efforts. For example, he can avoid social opprobrium and perhaps ostracism by not engaging in criminal activities; achieve distinction by working diligently at his occupation, giving to charities, or having a beautiful house; or relieve his envy and jealousy by talking meanly about or even physically harming his neighbors. These effects can be formalized in a production function for the $(R_j^1, \ldots, R_j^r)$ that depends partly on the efforts of i and partly on other variables.

---

[7] For an exposition of this theory, see Michael and Becker (1973).

To simplify the discussion,[8] I follow Senior and assume only a single commodity (distinction?) that is produced with a single good (the input of time is ignored) and a single characteristic of others. Then maximizing utility is equivalent to maximizing the output of this commodity, and one can write

$$U_i = Z(x, R). \tag{2.3}$$

I assume also (until Section 3C) that the effect of other variables (including the efforts of others) on this characteristic is not dependent on i's own efforts. Therefore, R can be written as the additive function

$$R = D_i + h, \tag{2.4}$$

where h measures the effect of i's efforts, and D_i the level of R when i makes no effort; that is, D_i measures i's "social environment."

His budget constraint for money income can be written as

$$p_x x + p_R h = I_i, \tag{2.5}$$

where I_i is his money income, $p_R h$ is the amount he spends on R, the price to him of a unit of R. Substitute $R - D_i$ for h in equation to get

$$p_x x + p_R R = I_i + p_R D_i = S_i.$$

The right-hand side gives the sum of i's money income and the value him of his social environment, and will be called his social income. The left-hand side shows how his social income is "spent": partly on his "own" goods (x) and partly on the characteristics of others (R).[9]

If i maximizes the utility-output function given by equation (2.3) subject to the constraint on social income given by equation (2.6), the equilibrium condition is[10]

$$\frac{\partial U_i}{\partial x} \bigg/ \frac{\partial U_i}{\partial R} = \frac{p_x}{p_R}. \tag{2.7}$$

---

[8] I have also developed the analysis assuming many commodities and many characteristics.

[9] Sociologists sometimes assert that variables like social approval and respect "do not have any material value on which a price can be put" (see Blau 1968). But prices measure only scarcity and have nothing intrinsically to do with "material value"; p_R, for example, only measures the resource cost to i of changing social approval, respect, etc.

[10] I assume for simplicity in this formula that p_R measures the marginal as well as average price of R.

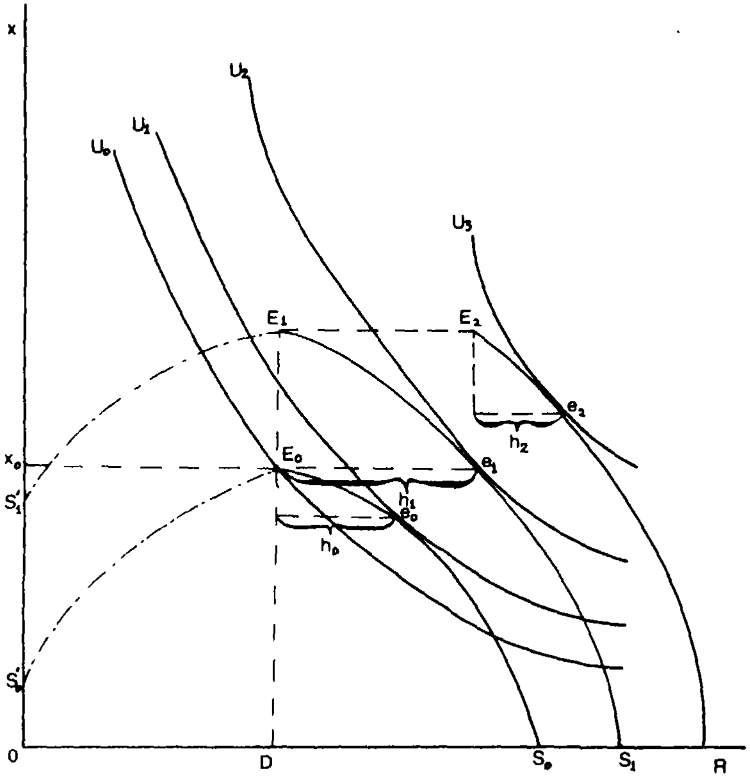


FIG. 1

If I did not want to purchase any R, p_R would be a "shadow" price, measured by the monetary equivalent of the marginal utility (equal to the marginal product) of R to i when $R = D_i$ (or when $h = 0$).

His equilibrium position is shown in figures 1 and 2. The first figure assumes that R has a positive marginal product in the production of Z (a positive marginal utility); that R refers, for example, to the respect accorded i rather than to his envy of others. The quantity $0D$ measures his social environment, and $0x_0$ his own income (measured in terms of x), so that the "endowed" point E_0 gives his utility when he spends nothing on R. If E_0S_0 measures the opportunities available for purchasing additional R,[11] he would maximize his utility by moving along E_0S_0 to point e_0, where the slope of this opportunity curve equaled the slope of his

---

[11] If he can also reduce R by giving up own goods, the curve E_0S_0 would continue in the southwest direction (see ES_0' in the figure). However, this section would be irrelevant if R had positive marginal utility.

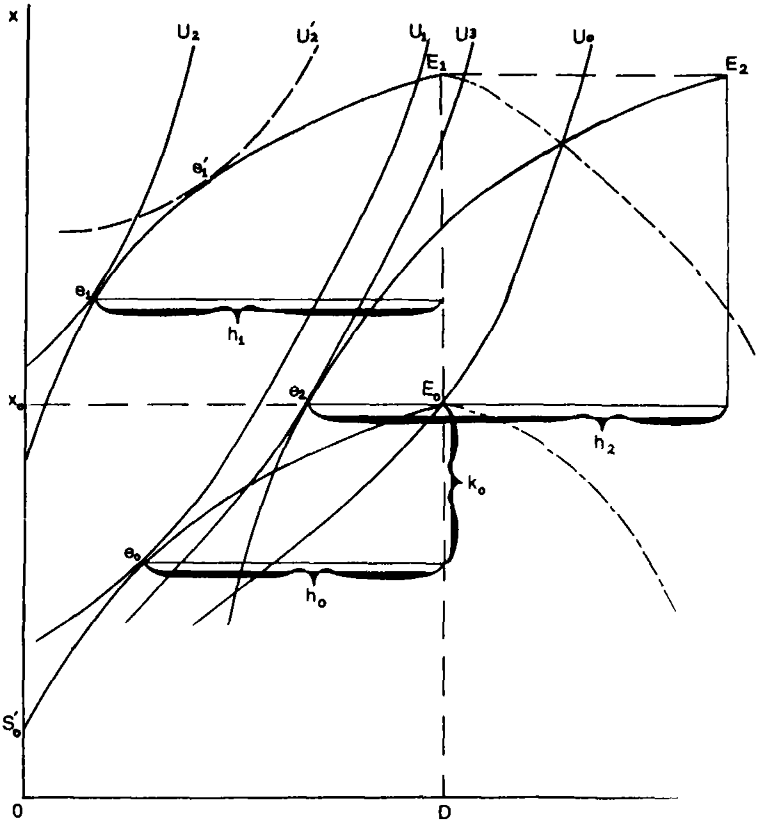


FIG. 2

indifference curve. His equilibrium purchase of R is measured by the line segment h_0.

Figure 2 assumes that R has a negative marginal product (or utility) because, say, it measures the income or prestige of persons that i envies. The section of the opportunity curve to the southeast of point E_0 is now irrelevant, and he moves along the southwest section E_0S_0' to point e_0. He is willing to give up resources to reduce R because his utility is raised by a reduction in R; at point e_0, he spends enough resources to reduce R by h_0.

Note that since the marginal (and average) price of R is negative in figure 2, i's social income is *less* than his own income because the value of his social environment is subtracted from his own income. That is, he is made worse off by his social environment if it is dominated by characteristics of others that are distasteful to him. Note too that as long as the marginal utility of R is not zero at the socially endowed position, his social income would differ from his own income even if he did not want to spend anything on R. He would add to (or subtract from) his own income the product of D and the (monetary equivalent of the) marginal utility of R at the endowed position E_o. In other words, the traditional income concept is incomplete even when no resources are spent trying to influence the attitudes or situation of others.

The analysis developed for social interactions in these figures and in equations (2.3), (2.6), and (2.7) is also applicable whenever there is a physical environment that either can be altered directly or can have its effects augmented or diminished. For example, the human capital of a person is the sum of the amount inherited and that acquired through investments; moreover, the amount invested is partly determined by the inheritance. Or the temperature in a house is determined by the weather and expenditures on fuels, insulation, etc., that reinforce or offset the natural environment.

A more general analysis, therefore, would assume that every term entering the utility function has both an environmental and acquired component. The general analysis could readily be developed, but I have chosen to simplify the discussion by ignoring the nonsocial environment. The results are consistent with those from the general analysis as long as the contribution of the social environment is, on the whole, significantly more important than that of the physical environment. This is assumed to be true. (I am indebted to Gilbert Ghez and especially Robert Barro for stressing the general nature of the analysis.)

B. *Income and Price Effects*

An increase in i's own income alone—without any change in prices or the social environment—would increase both x and R unless one were inferior. The average percentage response in x and R per 1 percent change in his own income is not unity, but is less by the fraction α, where α is the share of the social environment in his social income.[12] Therefore, the effect of a change in his own income on his utility-output is smaller the more important his social environment is.

Put differently, the greater the contribution of his social environment to his social income, the more his welfare is determined by the attitudes and behavior of others rather than by his own income. Traditional models

---

[12] By differentiating equation (2.6) with respect to I_i alone, $\bar{n} \equiv w_x n_x + w_R n_R = 1 - \alpha$, where

$$w_x = \frac{p_x x}{S_i}, w_R = \frac{p_R R}{S_i} = 1 - w_x, n_x = \frac{dx}{dI_i} \cdot \frac{I_i}{x}, n_R = \frac{dR}{dI_i} \cdot \frac{I_i}{R}, \alpha = \frac{p_R D_i}{S_i},$$

and I am assuming that p_R is given (not dependent on h, x, etc.). Of course, the weighted average of income elasticities with respect to a change in S_i must equal unity, as in the usual analysis.

of choice by economists assume that own efforts and access to property income and transfer payments determine welfare. On the other hand, those who stress the social environment, its normative requirements and sanctions for compliance and noncompliance, and the helplessness of the individual in the face of his environment naturally see society dominating individual efforts and, consequently, see little scope for important choices by individuals.

The relative importance of the social environment, as well as other implications of the theory of social interactions, can be empirically estimated from information on expenditures motivated by these interactions. If i's social environment did not change when his own income changed, the induced absolute change in the characteristics of others would equal the change in his contribution to these characteristics. However, the relative change in his contribution would differ from the relative change in these characteristics because the level of the latter is partly determined by the social environment.

Consider again figures 1 and 2, where an increase in i's own income with no change in the environment is shown by a vertical increase in the endowed position from E_o to E_1. Since his equilibrium position changes from e_o to e_1, the change in R is exactly equal to $h_1 - h_o$, the change in i's contribution to R. The percentage change in R in figure 1 is than that in h, since R is the sum of h and (a fixed) D. Since the change in R in figure 2 is negative, it is also less than the change in h, which is positive (since h is negative). However, if R increased by the increase in i's own income—if, say, the new position was at point e_1'—the percentage change in R would and would clearly exceed in algebraic value the negative change in h.

The own-income elasticity of demand for contributions is related to the elasticity of demand for characteristics by the following formula:[13]

$$n_h \equiv \frac{dh}{dI_i} \cdot \frac{I_i}{h} = \frac{n_R}{\bar{n}(=1-\alpha)}\left[1 + \alpha\left(\frac{1}{\beta} - 1\right)\right], \tag{2.8}$$

where $0 \leq \beta \leq 1$ is the fraction of own income that is spent on contributions to R. If $\alpha > 0$, if the social environment adds to i's social income,

---

[13] Since $dh/dI_i = dR/dI_i$,

$$n_h = \frac{dh}{dI_i} \cdot \frac{I_i}{h} = \frac{dR}{dI_i} \cdot \frac{I_i}{R} \cdot \frac{R}{h} = n_R \cdot \frac{R}{h}. \tag{2.8'}$$

But

$$\frac{R}{h} = \frac{p_R R}{p_R h} = 1 + \frac{p_R D_i}{p_R h} = 1 + \frac{S_i - I_i}{\beta I_i} = 1 + \frac{1/(1-\alpha) - 1}{\beta} = \frac{(1-\alpha) + \alpha/\beta}{1-\alpha}.$$

Since $1 - \alpha = \bar{n}$ (see n. 12 above), $n_h = (n_R/\bar{n})(\alpha/\beta + 1 - \alpha)$.

then clearly $n_h > n_R$.[14] Moreover, if $n_R \geq \bar{n} = 1 - \alpha < 1$; necessarily $n_h > 1$ even when $n_R < 1$; that is, contributions to the characteristics of others could have a "high" income elasticity even when the characteristics themselves had a "low" elasticity. Of course, if $n_h > 1$, the own-income elasticity of demand for own consumption (n_x) would be less than unity. That is, social interaction implies a relatively *low* income elasticity for own consumption even without introducing transitory changes in income, errors in variables, and the like.

Equation (2.8) further implies that an increase in α, an increase in the social environment, with no change in the own-income elasticity of demand for characteristics relative to the average elasticity $(n_R/\bar{n})$,[15] would increase the own-income elasticity of demand for contributions.[16] In other words, the more that *i*'s social income was determined by his social environment, the greater would be the percentage change in his contributions to the characteristics of others as his own income changed.

If, on the other hand, $\alpha < 0$—the social environment subtracted from *i*'s social income—then equation (2.8) implies that $n_h < n_R$ when $n_R > 0$, and $n_h > n_R$ when $n_R < 0$ (these different cases are shown in fig. 2). His demand for characteristics would probably be reduced by an increase in his own income (i.e., $n_R < 0$) if these characteristics have a negative marginal utility to him. Again, an increase in α, with $n_R/\bar{n}$ held constant, would raise n_h (the argument in n. 16 fully applies).

Since the social environment to any person cannot be readily observed, an indirect method of estimating at least its sign would be useful. If $n_R/\bar{n}$ were known, that is, if the relative income elasticity of demand for characteristics were known, the sign of α could be estimated simply from information on the own-income elasticity of demand for contributions to the environment, and its magnitude from additional information on the fraction of own income spent on these contributions. Equation (2.8) implies that

$$\alpha = \frac{n_h(\bar{n}/n_R) - 1}{1/\beta - 1}. \tag{2.9}$$

Therefore, $\alpha \gtreqless 0$ as $n_h(\bar{n}/n_R) \gtreqless 1$, and information on n_h, $\bar{n}/n_R$, and β would be sufficient to estimate α.

An increase in a social environment that adds to *i*'s social income would increase his demand for own goods if they had positive income elasticities.

---

[14] For $[1 + \alpha(1/\beta - 1)]/(1 - \alpha) > 1$, since $1/\beta > 1$, and $1 - \alpha < 1$.

[15] An increase in α lowers $\bar{n}$ because the relative contribution of own income to social income is reduced.

[16]
$$\frac{dn_h}{d\alpha}\left(\frac{n_R}{\bar{n}} = \text{constant}\right) = \frac{n_R}{\bar{n}}\left(\frac{1}{\beta} - 1\right) - \frac{n_R}{\bar{n}}\,\alpha\,\beta^{-2}\,\frac{d\beta}{d\alpha}.$$

Both terms are greater than zero because $\beta < 1$, and $d\beta/d\alpha < 0$ (this is shown shortly); therefore, $dn_h/d\alpha > 0$.

If his own income were unchanged, his increased expenditure on own goods would have to be "financed" by reduced contributions to the characteristics of others. Similarly, an increase in a social environment that subtracts from his social income would increase his expenditures on others and reduce his expenditures on own goods. Consequently, the effect of a change in the environment is always (i.e., as long as own goods are not inferior) partly offset by induced changes in i's contributions in the opposite direction, regardless of whether the environment adds to or subtracts from i's social income.

Geometrically, a change in the social envirnment is shown by a horizontal movement of the endowed position. An increase in the environment shifts the endowment in figure 1 from point E_1 to E_2; the equilibrium position is changed from point e_1 to a point on a higher indifference curve (e_2), and i's contribution declines from h_1 to h_2. In figure 2, the equilibrium is changed from point e_1 to a point on a lower indifference curve (e_2), and i's contribution increases from h_1 to h_2.[17]

If both the own and environment incomes of i changed, the effect would be a combination of those when each alone changed. For example, if both incomes increased, the effect on his contributions of the increase in the environment would at least partly offset the effect of the increase in his own income. In particular, if both incomes increased by the same pe age, the percentage change in contributions would be greater th to, or smaller than that percentage as his demand for chara exceeded, equaled, or was less than unity.

Through the assumption that p_R is constant, I have been ass effect, that expenditures and the social environment are perf tutes in producing characteristics of others. However, the qualit plications of this assumption can also be derived if they are simply substitutes for each other than for own consumption—if p_R rises as h rises, but not "too" rapidly. For example, a rise in the environment would reduce contributions, and a rise in own income would increase contributions by a relatively large percentage if the environment and expenditures on these characteristics are simply relatively close direct substitutes.

A rise in the cost of changing the characteristics of others (p_R) would induce the usual substitution (and perhaps income) effects away from these characteristics. If the environment were given, the absolute change in contributions would equal the absolute change in these characteristics,

---

[17] The endowment-income elasticity of demand for contributions can easily be shown to equal

$$N_h = \frac{dh}{dD} \cdot \frac{D}{h} = (N_R - 1) \left\{\frac{1}{1 - \alpha} \left[1 + \alpha \left(\frac{1}{\beta} - 1\right)\right]\right\} + 1.$$

Clearly, when $\alpha > 0$, $N_h < 0$ if $N_R \leq \alpha = \bar{N}$, the average endowment-income elasticity of demand; and when $\alpha < 0$, $N_h > 0$ if $N_R \geq \alpha$.

while the percentage changes would differ according to equation (2.8) in the following way:

$$E_h = -\frac{dh}{dp_R}\frac{p_R}{h} = E_R\left[\frac{1 + \alpha(1/\beta - 1)}{1 - \alpha}\right] \tag{2.10}$$

(same proof as in n. 13 above). Therefore, when $\alpha > 0$, E_h would exceed E_R by an amount that would be greater, the greater α and the smaller β. Similarly, when $\alpha < 0$, E_h would be less than E_R[18] by an amount that would be greater, the greater the absolute value of α and the smaller β.

3. Applications

Three specific applications of the general analysis of social interaction are now considered: interactions among members of the same family, charity, and envy and hatred. These applications not only provide empirical support for the income and price implications just derived, but also bring out a number of other implications of social interaction.

A. *The Family*

Assume that i cares about his spouse j in the sense that i's utility function depends on j's welfare.[19] I assume until much later in this section that j does not care positively or negatively about i. For simplicity, define the variable measuring this dependence, R_i, as follows:

$$R_i = \frac{I_j + h_{ij}}{p_x} = \frac{S_j}{p_x} = x_j , \tag{3.1}$$

where I_j is j's own income, h_{ij} are the contributions from i to j, S_j is j's social income ,and x_j are the goods consumed by j. The social income of i can be derived by substituting equation (3.1) into equation (2.6):

$$p_x x_i + p_R R_i = S_i = I_i + \frac{p_R I_j}{p_x} , \tag{3.2}$$

where p_R is the price to i of transferring resources to j. If i can transfer resources to j without any "transactions" costs—presumably, these costs are reduced by sharing a common household—and if i cares sufficiently about j to have $h_{ij} > 0$, then $p_R = p_x$, and

$$S_i = p_x x_i + p_x x_j = I_i + I_j = I_{ij} . \tag{3.3}$$

---

[18] I assume that an increase in the absolute value of p_R reduces the demand for R, so that $E_h > 0$.

[19] Caring is not simply a *deus ex machina* introduced to derive the following implications, since I have shown elsewhere (Becker 1974) that the marriage market is more likely to pair a person with someone he cares about than with an otherwise similar person that he does not care about.

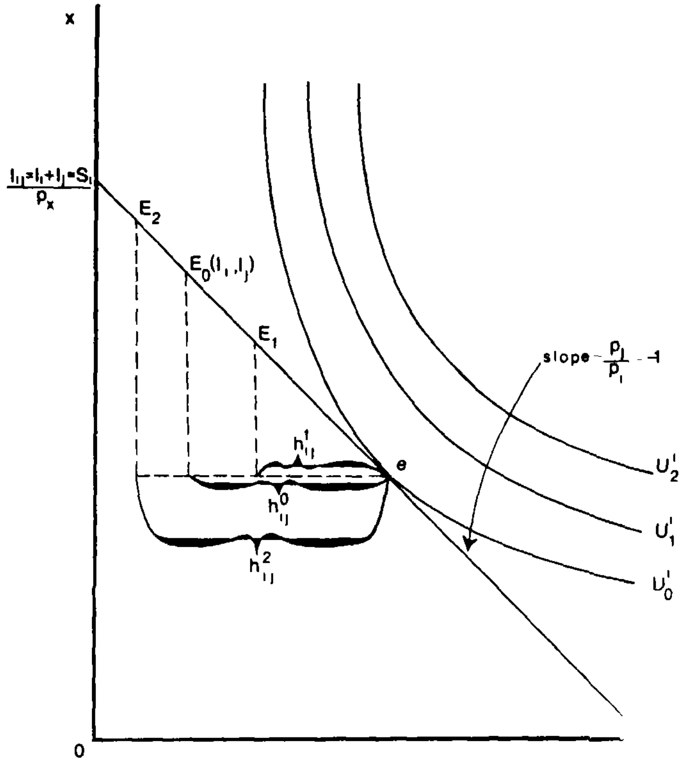


FIG. 3

The social income of i equals the combined own incomes of i and j, or the "family's" own income. Moreover, the equilibrium condition given by equation (2.7) implies that

$$\frac{\partial U_i}{\partial x_i} \Big/ \frac{\partial U_i}{\partial (R_i = x_j)} = \frac{p_x}{p_R} = 1\,, \tag{3.4}$$

or i would receive equal marginal utility from j's and his own consumption.

Conditions (3.3) and (3.4) are shown in figure 3. Resources can be transferred from i to j by moving along i's budget line in a southeast direction from the endowed position at point E_0. The equilibrium position is at point e, where the slope of i's indifference curves equals the slope of his budget line (= to $-$ 1). The vertical (or horizontal) intercept gives the family's own income—i's social income—deflated by the price of x.

An important implication of this analysis is that a change in the distribution of family income between i and j has no effect at all on the consumption or welfare of either, as long as i continues to transfer resources to j. A change in the distribution would be on the same budget line as E_0 if total family income is unchanged: the change from E_0 to E_1 is nominally more favorable to j, whereas the change to E_2 is nominally more favorable to i. Since there is only one point of tangency between i's budget line and an indifference curve, the equilibrium position must be unchanged at e. A shift in favor of j's income to E_1 simply induces an equal reduction in i's contributions to j (from h_{ij}^0 to h_{ij}^1 in the figure), whereas a shift against j's income to E_2 induces an equal increase in his contributions (from h_{ij}^0 to h_{ij}^2).[20]

This discussion has assumed a two-person family but is equally applicable to larger families that include grandparents, parents, children, uncles, aunts, or other kin. If one member, call him the "head," cares sufficiently about all other members to transfer general resources to them,[21] redistribution of income among members would not affect the consumption of any member, as long as the head continues to contribute to all.

The head's concern about the welfare of other members provides each, including the head, with some insurance against disasters. If a disaster reduced the income of one member alone, k, by say 50 percent, the head would increase his contributions to k, and thereby offset to some extent the decline in k's income. The head would "finance" his increased contribution to k by reducing his own consumption and his contributions to other members; in effect each member shares k's disaster by consuming less. If k's share of family income were negligible, he would essentially be fully insured against his own disasters because even a 50 percent decline in his income would have a negligible effect on family income, and thus on the consumption of each member. Since the share contributed by any member would tend to be inversely related to family size, large families, including the extended kinship family found in certain societies, can provide self-insurance especially when old-age, health, and other kinds of market insurance are not available or are very costly.[22] Note that insurance is automatically provided when resources are voluntarily transferred, without the need for any member to have dictatorial control over the family's allocation of resources.

---

[20] If the utility of i also partly depended directly on the amounts he transferred to j, perhaps because i's "prestige" or "approval" partly depended on these transfers, then redistribution of family income *would* have a net effect on the consumption of both i and j.

[21] A somewhat weaker assumption is that the family is "fully connected" through a series of transfers between members; for example, a transfers resources to b because a cares about b, b transfers to c because b cares about c, and so on until m transfers to the last member, n, and n transfers to no one (this assumption is made in an intergenerational context by Barro 1974). Indirectly, a (or any other member but n) would be transferring to all members because an increase in his contributions to b would induce an increase in the contributions to all other members.

[22] The interaction between self and market insurance is analyzed in Ehrlich and Becker (1972).

The result on the unimportance of the distribution of income among persons linked by transfers can also be used to understand the interaction among generations.[23] Suppose that the resources of the present generation are changed at the expense of or to the benefit of the resources accruing to future generations. For example, increased government debt or social security payments are financed by increased taxes on future generations, or increased public investment, perhaps in schools, with benefits accruing to future generations is financed by taxes on the present generation. If present and future generations are fully connected by a series of intergenerational transfers, called "bequests," then each of these apparent changes in the relative resources of present and future generations would tend to be offset by equal but opposite changes in bequests. In particular, increased public debt would not raise the real wealth or consumption of the present generation or reduce that of future generations because increased taxes on future generations would be matched by increased bequests to them. Similarly, increased public investment in education would be matched by reduced private investment in education.[24]

The budget constraint of the head is determined by total family income, not his own income alone—equation (3.3) for a two-person fam
be readily generalized to many persons. Since the head maxi
utility subject to his budget constraint, anything that increas
income would increase his utility. Therefore, the head would co
effect on total family income of his different actions, and would fo
income if the incomes of other family members were increased ev
For example, he would not move to another city if his spouse's o
ren's income would be decreased by more than his own income would b
increased. Or, although children usually eventually set up their own households and fully control their own incomes, the head would guide and help finance their investments in education and other human capital to maximize the present value of the real income yielded by these investments.[25]

Put differently, the head automatically internalizes the "external" effects of his actions on other family members.[26] Indeed, because the

[23] This application is taken from the detailed discussion in Barro (1974).

[24] The empirical evidence does strongly suggest that most of the investment in higher education by state governments has been offset by reduced private investment (see Peltzman 1973; McPherson, in preparation).

[25] The incentive that parents have to invest in their children is discussed in several places (see, e.g., De Tray 1973; Parsons 1974).

[26] The Coase Theorem proves that when "bargaining costs" are negligible, each family member could always be induced to maximize family opportunities through bargaining with and side payments from other members. I have proved that the head (and, as shown later, other members too) has this incentive and, in effect, makes or receives "side payments" without bargaining with other members. The word "automatically" is used to distinguish this theorem from the Coase Theorem.

head maximizes family income, he *fully* internalizes these externalities not only when the income of different members but also when their consumption, the other side of the budget constraint, is directly affected. He would take an action directly affecting consumption only when either the value of any increase in his consumption exceeded the value (to him) of any decrease in other members' consumption, or when any decrease in his own was less valuable than the increase in theirs.[27]

For example, he would read in bed at night only if the value of reading exceeded the value (to him) of the loss in sleep suffered by his wife, or he would eat with his fingers only if its value exceeded the value (to him) of the disgust experienced by his family. The development of manners and other personal behavior "rules" between family members well illustrates how apparent "external" effects can be internalized by social interaction between members.

Note too that not only is the head better off when his utility is raised, but so too are other members of his family, even if his actions directly reduce their consumption or increase their discomfort and disgust. For if his utility is raised and if their welfare has a positive income elasticity to him, he would increase his contributions to them by more than enough to offset their initial losses. For example, if he benefits from reading at night, his wife does too because he more than compensates her for her loss of sleep.[28]

The head maximizes a utility function that depends on the consumption of all family members subject to a budget constraint determined by family

---

[27] Although this is a rather immediate implication of his interest in maximizing family opportunities, a direct proof may be instructive. Suppose that a particular action changed the utility of the head by

$$dU_h = mu^h dx_h + \sum_{j=1, \neq h}^{n} mu^j dx_j, \tag{1'}$$

where $mu^j = \partial U_h / \partial x_j$, and dx_j measures the change in consumption of the jth family member. If the head can transfer resources to other members dollar for dollar, in equilibrium,

$$mu^j = \lambda_h p_j \qquad \text{all } j, \tag{2'}$$

where λ_h is the marginal utility of income to the head, and p_j is the cost of x_j. Substitution of eq. (2′) into (1′) gives

$$dU^h = \lambda_h \Big(p_h dx_h + \sum_{j=1, \neq h}^{n} p_j dx_j\Big) = \lambda_h \sum_{\text{all } j} p_j dx_j. \tag{3'}$$

Since the head takes an action if and only if $dU_h > 0$, eq. (3′) implies (since $\lambda_h > 0$) that he takes an action if, and only if,

$$\sum_{\text{all } j} p_j dx_j > 0, \tag{4'}$$

which was to be proved.

[28] Recall that I have been assuming that only a single good is consumed by each person, although this analysis presupposes many goods. The transition to many goods is straightforward if the head's utility depends on a function of the various goods consumed by another member that is monotonically related to the utility function of that member (see the discussion later in this section).

income and family consumption. Therefore, the effect of a change in relative prices of goods, or in aggregate family income (as well as in its distribution) on a family's consumption of different goods, could be predicted solely from the head's utility function and a budget constraint on family variables. The usual substitution and income effects of demand theory would be fully applicable.

In this sense, then, a family with a head can be said to maximize "its" consistent and transitive utility function of the consumption of different members subject to a budget constraint defined on family variables. The "family's" utility function is identical with that of one member, the head, because his concern for the welfare of other members, so to speak, integrates all the members' utility functions into one consistent "family" function.

That is, a "family's" utility function is the same as that of one of its members not because this member has dictatorial power over other members, but because he (or she!) cares sufficiently about all other members to transfer resources voluntarily to them. Each member can have complete freedom of action; indeed, the person making the transfers would not change the consumption of any member even with dictatorial power! *For example, if i had dictatorial power, he could move the equilibrium* position *e* in figure 3 to the vertical axis (or anywhere else), but would not choose to move it because his utility partly depend consumption.[29]

Nothing much has yet been said about the preferences of mem are not heads. The major, and somewhat unexpected, conclusion is that if a head exists, *other members also are motivated to maximize family income and consumption, even if their welfare depends on their own consumption alone.* This is the "rotten kid" theorem (I owe this name to the Barro family). For consider a selfish member j who can take an action that would reduce his income by b, but increase that of another member k by c. Initially, j would be worse off by b, since the gain to k is of no direct concern to him. However, if $c = b$, the head would transfer enough additional resources to j from k to leave him (and k) equally well off, since intrafamily reallocations of income do not affect the consumption of any member. Moreover, if $c > b$—if family income were raised by j's action—and if j's welfare were a superior "good" to the head, then he would transfer enough additional resources to j to make j better off. Consequently, even a selfish j would only undertake actions that raised family income or consumption, regardless of the initial impact on him.

---

[29] It is difficult to contrast my derivation of a "family" utility function w ditional derivation, since explicit derivations are rare. The most explicit appear a well-known article on social indifference curves by Samuelson (1956). He considers problem of relating individual and family utility functions, but his discussion is brief and the arguments sometimes are not spelled out. Without sufficient elaboration, he refers to a *consistent "family welfare function" being grafted onto the separate utility functions of* different family members (p. 10). In addition, he says that a family member's "preferences among his own goods have the special property of being independent of the other members' consumption. But since blood is thicker than water, the preferences of the different members are interrelated by what might be called a 'consensus' or 'social welfare function' which takes into account the deservingness or ethical worths of the consumption levels of each of the members." How are these preferences interrelated by a "*consensus*," and should *not* the "deservingness" of the consumption levels of different members simply be incorporated into different members' preferences (as in my approach)? Incidentally, at one point (p. 9), Samuelson appears to believe that if the family utility function is the same as the head's, he must have sovereign power, which I have shown is not necessary. He later (p. 20) says that "if within the family there can be assumed to take place an optimal reallocation of income so as to keep each member's dollar expenditure of equal ethical worth, then there can be derived for the whole family a set of well-behaved indifference contours relating the totals of what it consumes: the family can be said to *act as if* it maximizes such a group preference function" (italics in original). In my analyses, the "optimal reallocation" results from interdependent preferences and voluntary contributions, and the "group preference function" is identical with that of the "head."

In other words, when one member cares sufficiently about other members to be the head, all members have the same motivation as the head to maximize family opportunities and to internalize fully all within-family "externalities," regardless of how selfish (or, indeed, how envious) these members are. Even a selfish child receiving transfers from his parents would *automatically* consider the effects of his actions on other siblings as well as on his parents. Put still differently, sufficient "love" by one member guarantees that all members act as if they loved other members as much as themselves. As it were, the amount of "love" required in a family is economized: sufficient "love" by one member leads all other members by "an invisible hand" to act as if they too loved everyone.

Armed with this theorem, I do not need to dwell on the preferences of nonheads. Of course, just as there may be no head if all members are sufficiently selfish, so there may be none if they are all sufficiently altruistic. Each would want to transfer resources to other members, but no one would want to accept transfers. Aside from that, mutual interaction or mutual interdependence of welfare raises no particular problems.[30]

---

[30] It frequently has been alleged to me that mutual interaction of the form

$$U_i = U_i[x_i, g_i(U_j)]$$

$$U_j = U_j[x_j, g_j(U_i)],$$

where x_i and x_j are the own consumption of i and j, and g_i and g_j are monotonic functions of the utility indexes U_i and U_j, results in instability and unbounded utility levels. For it is argued, an increase in x_i by one unit directly raises i's utility, which raises j's utility through g_j, which in turn further raises i's utility, and so on, until U_i and U_j approach infinity. Mathematically, there is an infinite regress, since, by substitution,

$$U_i = U_i[x_i, g_i\{x_j, g_j\{x_i, g_i\{x_j, g_j\{\dots\}] .$$

However, with appropriate restrictions on the magnitude of the interactions, the infinite regress has a finite effect, and the "reduced forms" of U_i and U_j on x_i and x_j

By assuming in figure 3 and in the formal development given by equations (3.1)–(3.4) that only a single good is consumed by each person, I eliminated any distinction between transferring general purchasing power and transferring particular goods to another member. If each member consumes many goods, the conclusions in this section about family utility functions, internalization of within-family externalities, and so on fully hold only if the head is content to transfer general purchasing power. He would transfer in this form if his utility function depended on the utility of other members—that is, if his utility function could be written in the form

$$U_h = U_h[x_{h1}, \ldots, x_{hm}, g_1(x_{21}, \ldots, x_{2m}), \ldots, g_n(x_{n1}, \ldots, x_{nm})], \quad (3.5)$$

where x_{ij} is the quantity of the jth good consumed by the ith person, and

$$dg_i = 0 \left(= \sum_{j=1}^{m} \frac{\partial g_i}{\partial x_{ij}} dx_{ij} \right)$$

implies that the utility of the ith person is unchanged. If he is concerned not about the utility of other members but about their consumption of particular "merit" goods, the conclusions can be quite different. The systematic discussion of merit goods is postponed to Section 3C.

If parents are transferring resources to their children in th
of gifts and expenditures on education and other human ca
they die in the form of bequests, then an increase in the incom
by a given percentage would tend to increase contributions to
a still larger percentage, certainly by one exceeding the incre
of their children (see the discussion in Section 2). In other w
butions to children can be very responsive to a change in parent
without the welfare of children being so responsive.

Empirical evidence on bequests, gifts, and many other transfers to children is seriously deficient. The general impression is, however, that

---

are well defined. Consider, for example, the Cobb Douglas functions

$$U_i = x_i^{a_i} U_j^{b_i}$$
$$U_j = x_j^{a_j} U_i^{b_j},$$

where a_i and a_j presumably are greater than zero, and b_i and b_j can either be greater than or less than zero. By substitution,

$$U_i = x_i^{a_i/(1-b_ib_j)} x_j^{a_jb_i)/(1-b_ib_j)} = x_i^{\alpha_i} x_j^{\beta_i}$$
$$U_j = x_i^{a_ib_j)/(1-b_ib_j)} x_j^{a_j/(1-b_ib_j)} = x_i^{\alpha_j} x_j^{\beta_j},$$

where b_ib_j is independent of monotonic transformations on U_i and U_j. A finite sum to the regress requires that $|b_ib_j| < 1$; essentially, that the marginal utilities or disutilities due to interdependence are less than unity. Note that although it is possible for $a_i = b_i$ and $a_j = b_j$, for own consumption and the welfare of the other person to be equally "important," the condition $|b_ib_j| < 1$ implies that either $|\alpha_i| > |\beta_i|$, or $|\beta_j| > |\alpha_j|$, or both; that is, for at least one of the persons, own consumption has to be more important than the other person's consumption in the "reduced forms."

bequests have a very high income elasticity. Moreover, the elasticity of expenditures on children's education with respect to parental income does appear to be above unity (Schultz 1967, p. 9), which is consistent with the implications of the theory.

The responsiveness of expenditures on children's education and other training and skills to parental income has often been noted, and lamented as evidence of immobility and rigid "class" structure. Yet my analysis implies that the welfare of children—a measure of their "class"—rises by a smaller percentage than parent expenditures on them, and possibly even by a smaller percentage than parental income. Put differently, considerable regression toward the mean across generations—that is, the expected income or other measure of the position of children would be much closer to the average position than is that of their parents—can be observed at the same time that contributions to children are very responsive to parental income.[31]

The crucial point is that considerable regression toward the mean across generations would occur partly because of genetic factors and luck if all parents spent an equal amount on their children. As a result of this, *and given interdependent preferences, higher-income parents tend to* spend considerably more on their children than lower-income ones. However, these expenditures would only tend to dampen but not eliminate the

---

[31] In one study, the elasticity of children's years of schooling with respect to parental income is a sizeable +1.2, at the same time that the elasticity of children's *income* with respect to parental income is only +0.3, or a 70 percent regression toward the mean (unpublished calculations by Jacob Mincer from the Eckland Sample). Note in this regard, however, that parents cannot easily prevent considerable regression toward the mean by investing in their children. For let the relation between the human capital invested in children and parental income be

$$S_c = a + b \log I_p + u,$$

where b is the elasticity of parental response, and u represents other determinants of S_c. According to the theory of investment in human capital (Mincer 1974; Becker 1975 [in press]),

$$\log I_c = \alpha + rS_c + v,$$

where r is the rate of return on human capital, and v represents other determinants of $\log I_c$. Then by substitution,

$$\log I_c = (\alpha + ra) + rb \log I_p + (ru + v).$$

Even if r were as large as 0.2, and b as large as 2.0, rb would only be 0.4: the regression toward the mean would be 60 percent. If $v = c \log I_p + v'$, where $1 - c$ measures the degree of "intrinsic" regression to the mean, then by substitution,

$$\log I_c = (\alpha + ra) + (rb + c) \log I_p + (ru + tv').$$

Since the analysis in the text implies that b would be positively related to $1 - c$ as parents try to offset the "intrinsic" regression, the "observed" regression to the mean,

$$1 - \gamma = 1 - (c + rb) = (1 - c) - rb,$$

may be only weakly related to and also is less than the "intrinsic" regression $1 - c$. I am indebted to discussions with Jacob Mincer on the issues sketchily covered in this footnote.

regression toward the mean. Therefore, the elastic response of contributions to children can give a very biased picture of the degree of immobility or inheritance of "class" position. Indeed, contributions would be more responsive to parental income the stronger are the basic forces producing mobility because parents attempt to offset these forces. In other words, an elastic response of contributions to parental income may be evidence of sizeable *mobility*![32]

B. *Charity*

If someone makes contributions of time or goods to unrelated persons or to organizations, he is said to be "charitable" or "philanthropic." The discussion of contributions within a family indicates that charitable behavior can be motivated by a desire to improve the general well-being of recipients.[33] Apparent "charitable" behavior can also be motivated by a desire to avoid the scorn of others or to receive social acclaim. Not much generality is sacrificed, however, by only considering charity motivated by a desire to improve well-being.[34]

The numerous implications about family behavior developed in the previous section fully apply to the synthetic "family" consisting of a charitable person *i* and all recipients of his charity. For example, no n
well-being would be affected by a redistribution of income among

---

[32] It is generally believed that the United States has a more mobile "ope
than European countries do; yet (admittedly crude) comparisons of o
mobility between fathers and sons do not reveal large differences between
States and several Western European countries (Lipset and Bendix 1959
analysis in this paper suggests that parents' contributions to their children's
and other training is more responsive to parental position in "open" societies, mo
responsive parental contributions are probably offsetting the greater "openness" of American society.

[33] *The Random House Dictionary of the English Language* (unabridged, 1967) defines charity as "the benevolent feeling, especially toward those in need or in disfavor."

[34] The utility function of a charitable person who desires to improve the general well-being of recipients can be written as

$$U_i = U_i\left[x_i, x_j\left(= \frac{I_j + h}{p_j}\right)\right],$$

where *h* is his charitalbe giving, x_j measures the well-being of recipients, and $\partial U_i/\partial I_j = \partial U_i/\partial h > 0$; that is, a unit increase in the own income of recipients has the same effect on the utility of a charitable person as a unit increase in his giving. The utility function of a person who makes "charitable" contributions to win social acclaim can be written as

$$U_i = U_i\left(x_i, \frac{I_j}{p_j}, \frac{h}{p_j}\right),$$

where still $\partial U_i/\partial h > 0$—an increase in his contributions would increase his acclaim—but now the sign of $\partial U_i/\partial I_j$ is not so obvious. If, however, contributions and the income of recipients were much closer substitutes for each other than for the own consumption of the contributor, which is plausible, then these utility functions have similar implications. Not much generality is sacrificed, therefore, by only considering charity motivated by a desire to improve the well-being of recipients.

long as *i* continued to give to all of them. For he would simply redistribute his giving until everyone losing income was fully compensated and everyone gaining was fully "taxed." Moreover, all members, not simply *i*, would try to maximize "family" opportunities and "family" consumption, instead of their own income or consumption alone. In addition, each member of a synthetic "family" is at least partly "insured" against catastrophes because all other members, in effect, would increase their giving to him until at least part of his loss were replaced. Therefore, charity is a form of self-insurance that is a substitute for market insurance and government transfers. Presumably, the rapid growth of these latter during the last 100 years discouraged the growth of charity.

According to the analysis in Section 2, an increase in the income of a charitable person would increase his charitable giving by a greater percentage than the increase in the well-being of recipients. Indeed, his income elasticity of demand for giving would exceed unity, possibly by a substantial amount, as long as his elasticity of demand for their well-being (which I will call his demand for charity) was not much below his average income elasticity. The available evidence on charitable giving clearly supports this implication of the theory: income elasticities estimated by Taussig (1965) from giving in different income classes in 1962 are all well above unity, ranging from a low of +1.3 in the under $25,000 class to a high of +3.1 in the $100,000–$200,000 class.[35]

A crucial implication of charitable giving in terms of social interaction between the giver and others is that an increase in the incomes of recipients would reduce giving. Therefore, an increase in the incomes of both recipients and givers should not increase giving by as much as an increase in the incomes of givers alone. These implications are tested and confirmed by Schwartz (1970), who analyzes aggregate time series on incomes and charitable giving in the United States between 1929 and 1966 and also compares his findings with the cross-sectional findings of Taussig (1965) reported above.[36]

The usual theory of consumer choice ignores social interactions, and would consider charitable giving simply as a "good" that enters the giver's utility function along with his other goods:

---

[35] These estimates are net of differences in tax rates. Note, however, that charitable giving is estimated from itemized deductions in personal income tax returns. Since only giving to (certain) institutions and not to individuals can be deducted, since many taxpayers, especially with lower incomes, do not itemize their deductions, and since others inflate their deductions, the response of tax-reported giving may not accurately describe the response of actual giving.

[36] Schwartz's study, like Taussig's, is based on personal income tax returns. Both studies also estimate the price elasticity of giving, where price is measured by one minus the marginal tax rate. Schwartz finds considerable response to price, elasticities generally exceeding −0.5, which is consistent with the implications of the theory of social interactions. Taussig, on the other hand, finds only a weak response to price; but Schwartz argues that Taussig's findings are biased downward.

$$U_i = U_i(x_i, h) , \tag{3.6}$$

where h measures the amount given by i, and x_i are the other goods that he consumes. This "conventional" approach does not imply that an increase in i's income would increase his giving by a particularly large percentage, or that an increase in the incomes of recipients would lower his giving. Therefore, considerable ad hocery would be required if the "conventional" approach were to explain the evidence on charitable giving that is more readily explained by an approach that incorporates social interactions.

These findings can be used to make very crude, but instructive, calculations of the share of recipient's own incomes in the social incomes of contributors. If the own-income elasticity of demand for giving is taken from Taussig as +2.0, the share of own income spent on giving as 0.04 (see Schwartz 1970, p. 1278), and the income elasticity of demand for charity as equal to the average income elasticity (actually, Schwartz's findings suggest that it may be lower than the average), then, according to equation (2.9), charity's share in social income would be $(2 - 1)/(1/0.04 - 1) \approxeq 0.4$. If the own-income elasticity of giving were taken as +3.0 rather than +2.0, charity's share would double to 0.08
addition, the income elasticity of charity were only four-fift
average elasticity, its share would increase further to 0.11 (a tith

C. *Merit Goods and Multiperson Interactions*

Contributors are content to transfer general purchasing power
ients if they are concerned about the general welfare or utility of rec
ients—as seen by recipients. They want to restrict or earmark their transfers, on the other hand, if they are concerned about particular "merit" goods consumed by recipients. For example, parents may want transfers to their children spent on education or housing, or only the money incomes rather than "full" incomes of children may be of concern to parents, or contributors to beggars may not want their giving spent on liquor or gambling.

Assume, therefore, that i transfers resources to j that are earmarked for particular goods consumed by j because the utility function of i depends not only on his own goods but also on these goods of j. If j were permitted to spend his own income as he wished, an assumption modified shortly, he would spend less on these goods as a result of the earmarked transfers from i. Clearly, the reduction in his own spending would be greater, the greater the transfer, the smaller the fraction of his social income spent on these goods, and the smaller their income elasticity. For example, if they take 20 percent of his social income and have an income elasticity equal to 2.0, he would reduce his own spending by $0.60 for each dollar earmarked by i.[37]

As long as j continues to spend on the merit goods, earmarked transfers are worth as much to j as a transfer of general purchasing power with equal monetary value. Moreover, i would not have a greater effect on j's consumption of these goods with earmarked transfers than with general transfers. Therefore, as long as j continues to spend on these goods, earmarked transfers are equivalent to general transfers; and the results derived for the latter fully hold for the former. For example, a redistribution of income between i and j would have no effect on the consumption of either as long as both continue to spend on the merit goods, or both i and j want to maximize their combined incomes, not their own incomes alone.

On the other hand, if j did not want to spend anything on the merit goods because earmarked transfers were sufficiently large, such transfers would be worth less to j and more to i than would general transfers with equal money value. Moreover, various results derived for general transfers no longer hold: for example, a redistribution of income to j and away from i would reduce j's consumption of merit goods and increase his consumption of other goods.

If i were aware that j reduced his spending on merit goods when transfers increased, i would be discouraged from giving because j's reaction raises i's private price of merit goods to

$$p_m^i = p_m \frac{1}{1 - r_j} = p_m \frac{1}{v_m n_m},^{38} \tag{3.7}$$

where p_m is the market price of merit goods, and the other terms are defined in note 37. Similarly, if j were aware that i reduced his transfers when j increased his spending on merit goods, j would also be discouraged from spending because i's reaction raises the price to j. Indeed, j could end up consuming fewer merit goods than he would if i were not concerned! That these induced reactions are not simply hypothetical or always minor is persuasively shown in a recent study of higher education (Peltzman 1973). States earmark transfers to higher education mainly through highly subsidized public institutions. Private spending was apparently reduced by (at least) $0.75 per dollar of public spending in 1966–67; private spending may have been reduced by more than $1.00 per dollar of public spending in 1959–60, so that *total* spending on higher education in that year would have been reduced by public spending.

---

[37] It is easily shown that $r_j = 1 - v_m n_m$, where v_m is the share spent on merit goods; n_m, their income elasticity; and r_j, the reduction in j's own spending per unit increase in i's contribution. Therefore, if $v_m = 0.2$, and $n_m = 2.0$, $r_j = 0.6$.

[38] For example, if j spent $0.60 less for each dollar transferred by i, the price to i would be $p_m^i = p_m(1/0.4) = 2.5\,p_m$, or more than twice the market price.

Both i and j want to limit the induced reactions of the other because such reactions reflect the incentive to "underreveal" preferences about merit goods and "free-ride" in their consumption. Since equation (3.7) shows that these reactions raise the price of merit goods to i and j, in effect, both want to lower these prices. Indeed, it is well known from the theory of public goods, and a merit good is a particular kind of "public" good, that efficient prices to i and j would be less than the market price; indeed, these efficient prices would *sum* to the market price of the merit good.[39] Efficient prices might be achieved, for example, by i and j matching each other's spending in specified proportions, or each might be given a spending quota.

I intentionally say "might" be achieved because any agreement has to be "policed" to insure that each lives up to his commitment. Policing is relatively easy for the consumer of the merit goods, j, since he usually automatically knows how much is spent by i, but is much more difficult for i, since he does not automatically know how much is spent by j.[40] Parents may use their children's grades in school to measure the input of time and effort by children that presumably "matches" the money contribution by the parents.[41] Or parents may save a large part of their total transfer to children for a bequest when they die in order to provide an incentive for children to spend "appropriately," at least while parents are alive.[42] This may explain why the inheritance tax on apparently has induced relatively little substitution toward children (see Shoup 1966; Adams 1974).

The "underrevealing," "free-riding," coordination of eff "policing" discussed for merit goods are common to all multiper actions—that is, all situations where two or more persons are aff the consumption, attitudes, or other behavior of the same person. analytical issues for multiperson interactions are the same as for other "public" goods: is public intervention desirable—for example, should charitable giving be deductible from personal income in arriving at tax liabilities in order to lower the private price of giving—and do private equilibria without government intervention more closely approximate joint maximization, a Nash noncooperative game solution, or something quite different? Since space is limited, I refrain from discussing further these and related issues.

---

[39] A proof of this well-known summation formula can be found in Samuelson (1954).

[40] The difficulty of policing "merit" goods is shown amusingly in a recent Wizard of Id cartoon. Two drunks meet, and one says, "Could you spare a buck for a bottle of wine?" The other answers, "How do *I* know you won't buy food with it?"

[41] I owe this example to Lisa Landes.

[42] This conclusion about the incentives provided by large bequests is a special case of a more general result proven elsewhere (see Becker and Stigler 1974) that relatively large pensions discourage employees from acting contrary to the interests of their employers (a bequest serves the same purpose as a pension).

D. *Envy and Hatred*

An envious or malicious person presumably would feel better off if some other persons become worse off in certain respects. He could "harm" himself (i.e., spend his own resources) in order to harm others: in figure 2, he gives up k_0 units of his own consumption in order to harm others by h_0 units. The terms of trade between his own harm and the harm to others, given by the curve $E_0S_0^1$ in figure 2, is partly determined by his skill at "predatory" behavior and partly by public and private expenditures to prevent crime, libeling, malicious acts, trespass, and other predatory behavior. Since an increase in these expenditures would increase the cost to him of harming others, he would be discouraged from harming them. The limited evidence available on predatory expenditures supports this implication of the theory. Crimes against persons provide some evidence on predatory behavior, since most assaults and murders probably are motivated by the harm to victims.[43] The frequency of assaults and murder (and also crimes against property) apparently is strongly negatively related to the probability of conviction, punishment, and other measures of the cost of committing these crimes (see Ehrlich 1973).

Section 2 suggests that a rise in own income would tend to reduce predatory expenditures. An increase in the social environment,[44] on the other hand, would necessarily increase these expenditures, unless own consumption were an inferior good. Therefore, a rise in the social environment and own income by the same percentage would reduce predatory expenditures by less than would a rise in own income alone, and might even increase them.

Again, the implications of the theory can be tested with evidence on crimes against persons. Since assaults and murders have been more frequent at lower income levels,[45] an increase in own income appears to reduce crimes against persons, if differences in own income alone are measured by differences in the incomes of individuals at a moment in time (as in the discussion of charity in Section 3*B*). As predicted by the theory, an increase in own income that is accompanied by an increase in the social environment (as measured by the income of others) does not have such a negative effect on these crimes. Indeed, the frequency of assaults and murders has not been reduced by the sizeable growth in aggregate incomes during the last 40 years, nor do higher-income states presently have fewer crimes against persons than other states.[46]

---

[43] Most robberies, burglaries, and larcenies, on the other hand, probably are motivated by the prospects of material gain.

[44] That is, in that part of the social environment that motivates predatory expenditures.

[45] Persons committing crimes against other persons as well as against property are much more likely to live in low income areas (see Crime Commission 1967*a*, table 9).

[46] The rate of assaults grew significantly from 1933 to 1965 in the United States, and the murder rate remained about the same (Crime Commission 1967*b*, figs. 3, 4). Higher-

Over the years, even acute observers of society have differed radically in their assessment of the importance of envy and hatred. Two hundred years ago, for example, Adam Smith recognized these "passions" but shunted them aside with the comment: "Envy, malice, or resentment, are the only passions which can prompt one man to injure another in his person or reputation. But the *greater part of men are not very frequently under the influence of those passions,* and the very worst men are so only occasionally. As their gratification too, how agreeable soever it may be to certain characters, is not attended with any real or permanent advantage it is in the greater part of men commonly restrained by prudential considerations. Men may live together in society with some tolerable degree of security, though there is no civil magistrate to *protect them from the injustice of those passions*" (Smith 1937; my italics).[47] To Thorstein Veblen, on the other hand, writing many years later, there motives are the very stuff of life that dominate everything else: "The desire for wealth can scarcely be satiated in any individual instance, and evidently a satiation of the average or general desire for wealth is out of the question. However widely, or equally, or 'fouly,' it may be distributed, no general increase of the community's wealth can make any approach to satiating this need, the ground of which is the desire of everyone to excel everyone else in the accumulation of goods" (Veblen 1934, p.32).[48]

In principle, the importance of envy and hatred can be measur
equation (2.9) by the contribution of the relevant social enviro
social income; this is done in a crude way in Section 2*B* for
Unfortunately, not enough information is available either on
income elasticity of demand or on the fraction of own income
"predatory" behavior to make even crude estimates of the
contribution of envy and hatred.

Still, it may be useful to note several implications of the differing views about the significance of envy and hatred. For example, Veblen's belief that the welfare of a typical person primarily depends on his relative income position implies that social income essentially is zero: that the value of the social environment causing envy would exactly offset the value of own income.[49] For then, and only then, would a rise in this

---

income states do not have fewer crimes against persons even when the probability of conviction, the punishment, and several other variables are held constant (Ehrlich 1973, tables 2–5). Note that Ehrlich's study, unlike the evidence from the Crime Commission, holds the "price" of crime constant when estimating the effects of income (and holds income constant when estimating the effects of price).

[47] Not much later, Jeremy Bentham reached a similar conclusion: "The pleasure derivable by any person from the contemplation of pain suffered by another, is in no instance so great as the pain so suffered" (Bentham 1952–54).

[48] Similarly, a sociologist recently has argued that envy is a powerful motive in primitive as well as advanced societies, communist as well as capitalist ones, and is critical in determining economic progress and public policy (see Schoeck 1966).

[49] "Own" income here includes the value of other aspects of the social environment.

social environment and own income by the same percentage, prices held constant, not affect social income or welfare. That is, a rise in all incomes in a community by the same percentage would not improve anyone's welfare in Veblen's world.[50]

If social income were negative, if the environment causing envy were more important than own income, a rise in the environment and own income by the same percentage would lower social income and welfare. That is, a general rise in incomes in a more extreme Veblenian world would actually lower welfare![51]

On the other hand, Smith's belief that envy is a relatively minor determinant of welfare implies that social income is positive: the environment causing envy is less important than own income. A rise in the environment and own income by the same percentage would then raise social income and welfare. That is, Veblen's general rise in the community's income would raise the welfare of the typical person.

4. Summary

This essay uses simple tools of economic theory to analyze interactions between the behavior of some persons and different characteristics of other persons. Although these interactions are emphasized in the contemporary sociological and anthropological literature, and were considered the cornerstone of behavior by several prominent nineteenth-century economists, they have been largely ignored in the modern economic literature.

The central concept of the analysis is "social income," the sum of a person's own income (his earnings, etc.) and the monetary value to him of the relevant characteristics of others, which I call his social environment. The optimal expenditure of his own income to alter these characteristics is given by the usual marginal conditions. By using the concept of social income, I can analyze the effect on these expenditures of changes in different sources of income and in different prices, including the "price" of the social environment. Perhaps the most important implication is that a change in own income alone would tend to cause a relatively large change in these expenditures; in other words, the own-income elasticity of demand for these expenditures would tend to be "large," certainly larger than the elasticity resulting from equal percentage changes in own income and the social environment.

---

[50] If $U_i = U_i(I_i/\bar{I})$, where $\bar{I}$ is the average community income, then $S_i = I_i - p_r\bar{I}$, where S_i is i's social income, and p_r is the price of $\bar{I}$ in terms of I_i. If i did not engage in predatory behavior, p_r would simply equal the slope of his indifference curve: slope $= dI_i/d\bar{I} = I_i/\bar{I} = p_r$. Hence $S_i = I_i - I_i/\bar{I} \cdot \bar{I} = 0$.

[51] When envy is so important, economic development is undesirable because it lowers welfare. See Schoeck's (1966) discussion of what he calls "the envy-barrier of the developing countries."

Interactions among members of the same family receive the greatest attention. The "head" of a family is defined not by sex or age, but as that member, if there is one, who transfers general purchasing power to all other members because he cares about their welfare. A family with a head is a highly interdependent organization that has the following properties:

A redistribution of income among members does not affect the consumption or welfare of any member because it simply induces offsetting changes in transfers from the head. As a result, each member is at least partially insured against disasters that may strike him.

Not only the head but other members too act "as if" they "loved" all members, even when they are really selfish, in the sense that they maximize not their own income alone but family income. As it were, the existence of a head economizes on the amount of true love required in a family.

A family acts "as if" it maximized a consistent and transitive utility function subject to a budget constraint that depended only on family variables. This utility function is the same as the head's not because he has dictatorial power, but because his concern for the welfare of other members integrates all their utility functions into one consistent "family" function.

Transfers from parents to children in the form, say, of schooling, gift
and bequests tend to be negatively related to what the income of
would be relative to their parents in the absence of these
Therefore, the relative income of children *inclusive* of transfers
unrelated or even negatively related to these transfers. Conseque
cannot infer anything about the stability across generations of
or social positions simply from knowing the relation between
position and the amount transferred.

More briefly treated are charity and envy, with special attention the effects of different kinds of income change on charitable contributions and expenditures to alleviate envy. For example, the much higher income elasticity of demand for charitable contributions estimated from differences in individual incomes at a moment in time than from aggregate changes in incomes over time is shown to be implied by this theory of social interactions, but not readily by the traditional theory of choice.

From a methodological viewpoint, the aim of the paper is to show how another relation considered important in the sociological and anthropological literature can be usefully analyzed when incorporated into the framework provided by economic theory. Probably the main explanation for the neglect of social interactions by economists is neither analytical intractability nor a preoccupation with more important concepts, but excessive attention to formal developments during the last 70 years. As a consequence, even concepts considered to be important by earlier economists, such as social interactions, have been shunted aside.

References

Adams, James D. "Asset Transfers at Deathtime." Dept. Econ., Univ. Chicago, February 1974.

Alchian, A. A., and Allen, W. R. *University Economics.* 2d ed. Belmont, Calif.: Wadsworth, 1967.

Barro, R. "Are Government Bonds Net Wealth?" *J.P.E* 82, no. 6 (November/December): 1095–1117.

Becker, G. S. "Notes on an Economic Analysis of Philanthropy." Nat. Bur. Econ. Res., April 1961.

———. "Interdependent Preferences: Charity, Externalities, and Income Taxation." Univ. Chicago, March 1968.

———. *The Economics of Discrimination.* 2d ed. Chicago: Univ. Chicago Press, 1971.

———. "A Theory of Marriage: Part II." *J.P.E.* 82, no. 2, pt. 2 (March/April 1974): 11–26.

———. *Human Capital.* 2d ed. New York: Columbia Univ. Press, 1975, in press.

Becker, G. S., and Stigler, G. J. "Law Enforcement, Malfeasance, and Compensation of Enforcers." *J. Legal Studies* 3, no. 1 (January 1974): 1–18.

Bentham, J. *Principles of Morals and Legislation.* Oxford: Clarendon, 1789.

———. "The Philosophy of Economic Science." In *Jeremy Bentham's Economic Writings*, edited by W. Stark. 3 vols. New York: Franklin, 1952–54.

Blau, P. M. "Social Exchange." In *International Encyclopedia of the Social Sciences*, edited by D. E. Sills. Vol. 7. New York: Macmillan, 1968.

Boulding, K. *The Economy of Love and Fear.* Belmont, Calif.: Wadsworth, 1973.

Brady, D., and Friedman, R. D. "Savings and the Income Distribution." In *Studies in Income and Wealth*, Conference on Research in Income and Wealth. Vol. 10. New York: Nat. Bur. Econ. Res., 1947.

Crime Commission. *Crime and Its Impact: An Assessment.* Task Force Report. Washington: Government Printing Office, 1967. (*a*)

———. *The Challenge of Crime in a Free Society.* Task Force Report. Washington: Government Printing Office, 1967. (*b*)

DeTray, D. "Child Quality and the Demand for Children." *J.P.E.* 81, no. 2, pt. 2 (March/April 1973): 70–95.

Duesenberry, J. S. *Income, Savings, and the Theory of Consumer Behavior.* Cambridge, Mass.: Harvard Univ. Press, 1949.

Ehrlich, I. "Participation in Illegitimate Activities: A Theoretical and Empirical Investigation." *J.P.E.* 81, no. 3 (May/June 1973): 521–65.

Ehrlich, I., and Becker, G. S. "Market Insurance, Self-Insurance, and Self-Protection." *J.P.E.* 80, no. 4 (July/August 1972): 623–48.

Fisher, I. *Mathematical Investigations in the Theory of Value and Price.* New Haven, Conn.: Yale Univ. Press, 1926.

Johnson, H. "The Effect of Income-Redistribution on Aggregate Consumption with Interdependence of Consumers' Preferences." *Economica* (May 1952).

Leibenstein, H. "Bandwagon, Snob, and Veblen Effects in the Theory of Consumers' Demand." *Q.J.E.* 64 (May 1950): 183–207.

Lipset, S. M., and Bendix, R. *Social Mobility in Industrial Societies.* Berkeley: Univ. California Press, 1959.

McPherson, M. "The Effects of Public on Private College Enrollment." Ph.D. dissertation, Univ. Chicago, 1974.

Marshall, A. *Principles of Economics.* 8th ed. London: Macmillan, 1962.

Michael, R. T., and Becker, G. S. "On the New Theory of Consumer Behavior." *Swedish J. Econ.*, vol. 75, no. 4 (1973).

Mincer, J. *Schooling, Experience, and Earnings.* New York: Columbia Univ. Press, 1974.

Panteleoni, M. *Pure Economics.* Clifton, N.J.: Kelley, 1898.

Parsons, D. O. "Intergenerational Wealth Transfers and the Educational Decisions of Male Youth." Ohio State Univ., 1974.

Parsons, T. "Social Interactions." In *International Encyclopedia of the Social Sciences,* edited by D. S. Sills. Vol. 7. New York: Macmillan, 1968.

Peltzman, S. "The Effect of Government Subsidies-in-Kind on Private Expenditures: The Case of Higher Education." *J.P.E.* 81, no. 1 (January/February 1973): 1–27.

Pigou, A. C. "Some Remarks on Utility." *Econ. J.* 13 (1903): 19–24.

Samuelson, P. A. "The Pure Theory of Public Expenditures." *Rev. Econ. and Statis.* (November 1954).

———. "Social Indifference Curves." *Q.J.E.* (February 1956).

Schoeck, H. *Envy.* New York: Harcourt, Brace & World, 1966.

Schultz, T. W. *The Economic Value of Education.* New York: Columbia Univ. Press, 1967.

Schwartz, R. "Personal Philanthropic Contributions." *J.P.E.* 78, no. 6 (November/December 1970): 1264–91.

Shoup, C. *Federal Estate and Gift Taxes.* Washington: Brookings Inst., 1966.

Smith, A. *The Wealth of Nations.* New York: Modern Library, 1937.

Taussig, M. "The Charitable Contribution in the Federal Personal Income Tax." Ph.D. dissertation, Massachusetts Inst. Tech., 1965.

Veblen, T. *The Theory of the Leisure Class.* New York: Modern Library, 1934.

Vickery, W. S. "One Economist's View of Philanthropy." In *Philanthropy and Public Policy,* edited by F. Dickinson. New York: Nat. Bur. Econ. Res

Are Government Bonds Net Wealth?

Robert J. Barro

University of Chicago

The assumption that government bonds are perceived as net wealth by the private sector is crucial in demonstrating real effects of shifts in the stock of public debt. In particular, the standard effects of "expansionary" fiscal policy on aggregate demand hinge on this assumption. Government bonds will be perceived as net wealth only if their value exceeds the capitalized value of the implied stream of future tax liabilities. This paper considers the effects on bond values and tax capitalization of finite lives, imperfect private capital markets, a government monopoly in the production of bond "liquidity services," and uncertainty about fu tax obligations. It is shown within the context of an overlap generations model that finite lives will not be relevant to the capita tion of future tax liabilities so long as current generations connected to future generations by a chain of operative intergenera transfers (either in the direction from old to young or in the dir from young to old). Applications of this result to social security to other types of imposed intergenerational transfer schemes are al noted. In the presence of imperfect private capital markets, government debt issue will increase net wealth if the government is more efficient, at the margin, than the private market in carrying out the loan process. Similarly, if the government has monopoly power in the production of bond "liquidity services," then public debt issue will raise net wealth. Finally, the existence of uncertainty with respect to individual future tax liabilities implies that public debt issue may increase the overall risk contained in household balance sheets and thereby effectively reduce household wealth.

The assumption that government bonds are perceived as net wealth by the private sector plays an important role in theoretical analyses of monetary and fiscal effects. This assumption appears, explicitly or implicitly, in demonstrating real effects of a shift in the stock of public debt

I have benefited from comments on earlier drafts by Gary Becker, Benjamin Eden, Milton Friedman, Merton Miller, José Scheinkman, Jeremy Siegel, and Charles Upton. The National Science Foundation has supported this research.

[*Journal of Political Economy*, 1974, vol. 82, no. 6]

© 1974 by The University of Chicago. All rights reserved.

(see, e.g., Modigliani 1961, sec. IV; Mundell 1971; and Tobin 1971, chap. 5), and in establishing nonneutrality of changes in the stock of money (Metzler 1951, sec. VI). More generally, the assumption that government debt issue leads, at least in part, to an increase in the typical household's conception of its net wealth is crucial for demonstrating a positive effect on aggregate demand of "expansionary" fiscal policy, which is defined here as a substitution of debt for tax finance for a given level of government expenditure (see, e.g., Patinkin 1964, sec. XII.4; and Blinder and Solow 1973, pp. 324–25). The basic type of argument in a full-employment model is, following Modigliani (1961), that an increase in government debt implies an increase in perceived household wealth; hence, an increase in desired consumption (a component of aggregate demand) relative to saving; hence, an increase in interest rates; and, finally, a decline in the fraction of output which goes to capital accumulation. However, this line of reasoning hinges on the assumption that the increase in government debt leads to an increase in perceived household wealth. In a non-full employment context it remains true that the effect of public debt issue on aggregate demand (and, hence, on output and employment) hinges on the assumed increase in perceived household wealth.

It has been recognized for some time that the future taxes needed to finance government interest payments would imply an offset to the direct positive wealth effect. For example, in a paper originally published in 1952, Tobin (1971, p. 91) notes: "How is it possible that society merely by the device of incurring a debt to itself can deceive itself into believing that it is wealthier? Do not the additional taxes which are necessary to carry the interest charges reduce the value of other components of private wealth?" Bailey (1962, pp. 75–77) has gone somewhat further by arguing: "It is possible that households regard deficit financing as equivalent to taxation. The issue of a bond by the government to finance expenditures involves a liability for future interest payments and possible ultimate repayment of principal, and thus implies future taxes that would not be necessary if the expenditures were financed by current taxation. . . . If future tax liabilities implicit in deficit financing are accurately foreseen, the level at which total tax receipts are set is immaterial; the behavior of the community will be exactly the same as if the budget were continuously balanced."

There seem to be two major lines of argument that have been offered to defend the position that the offset of the future tax liabilities will be only partial.[1] One type of argument, based on finite lives, supposes that the relevant horizon for the future taxes (which might correspond to the remaining average lifetimes of the current taxpayers) will be shorter than that for the interest payments.[2] Accordingly, a stream of equal values for interest payments and taxes will have a net positive present value. This argument has been used explicitly by Thompson (1967, p. 1200). The second type of argument, usually based on imperfect private capital markets, supposes that the relevant discount rate for tax liabilities will be higher than that for the interest payments. Hence, even with an infinite horizon for tax liabilities, a stream of equal values for interest payments and taxes will have a net positive present value. This argument has been used by Mundell (1971).[3]

---

[1] Of course, most analyses of government debt effects do not offer a specific defense for this position. For example, Blinder and Solow (1973, p. 325, n. 8) say: "This [analysis] includes government bonds as a net asset to the public. We are well aware of, but not persuaded by, the arguments which hold that such bonds are not seen as net worth by individuals because of the implied future tax liability."

The first part of this paper deals with the effect of government bond issue on the calculus of individual wealth in an overlapping-generations economy with physical capital where individuals have finite lives. No elements of "capital market imperfections" are introduced into this model. The key result here is that, so long as there is an operative intergenerational transfer (in the sense of an interior solution for the amount of bequest or gift across generations), there will be no net-wealth effect and, hence, no effect on aggregate demand or on interest rates of a marginal change in government debt. This result does not hinge on current generations' weighing the consumption or utility of future generations in any sense on an equal basis with own consumption, nor does it depend on generations' placing any direct weight at all on the consumption of any future generation other than the immediate descendan generations act effectively as though they were infinite-lived w are connected to future generations by a chain of operat generational transfers.

The analysis then shows that social security payments are analogou changes in government debt. Marginal changes in this type (or oth types) of imposed intergenerational transfers have no real effects when current and future generations are already connected by a chain of operative discretionary transfers. The effects of inheritance taxes and of "transaction costs" for government bond issue and tax collections are also considered. It is shown that inheritance taxes do not affect the basic results, but that the presence of government transaction costs implies that the net-wealth effect of government bonds would actually be negative.

The second part of the paper deals with the existence of imperfect private capital markets. It is shown that, to the extent that public debt issue entails a loan from low-discount-rate to high-discount-rate individuals, a positive net-wealth effect results if the government is more efficient than the private market in carrying out this sort of loan. If the government is more efficient only over a certain range, and if the public choice process determines the amount of government debt issue in accord with efficiency criteria, it is again true at the margin that the net-wealth effect of government bond issue is nil.

---

[2] This type of argument applies to head taxes or to taxes based on wage income, but not to taxes which are based on the value of nonhuman assets. This distinction has been made by Mundell (1971, pp. 9, 10).

[3] A different line of argument that leads to a similar conclusion is that the government acts like a monopolist in the provision of the liquidity services yielded by its liabilities. I discuss this argument in part III, below.

The third part of the paper discusses government debt as a bearer of nonpecuniary "liquidity services." It is shown that if the government acts like a competitive producer of these services, as would be dictated by a public choice process which reflects efficiency criteria, then the net-wealth effect of government bond issue would be zero on this count. More generally, the net-wealth effect would be positive if the government acts like a monopolist and would be negative if the government is an overproducer of liquidity services.

The last part of the paper deals with the risk characteristics of government debt and of the tax liabilities associated with the interest payments on this debt. It is argued that if relative tax liabilities are known, a change in government debt will not alter the overall risk contained in household balance sheets. When relative tax liabilities are uncertain, the effect of government debt issue on the overall risk may be positive or negative, depending on the nature of the tax system and on the transaction costs associated with private insurance arrangements.

I. The Effect of Finite Lives—a Model with Overlapping Generation

A. Setup of the Model

I use here a version of the Samuelson (1958)-Diamond (1965) overlapping-generations model with physical capital. Each individual lives two periods, which will be distinguished by the superscripts y (young) and o (old). Generations are numbered consecutively beginning with the generation which is currently old (subscript 1); followed by its descendant, which is currently young (subscript 2); followed by its descendant; and so on. I assume here that there are the same number of people, N, in each generation, and that all individuals are identical in terms of tastes and productivity. I also abstract from any technological change over time. The members of each generation work (a fixed amount of time set equal to one unit) only while young and receive an amount of wage income w. Expectations on w for future periods (i.e., for future generations) are assumed to be static at the current value. Asset holdings (A) take the form of equity capital (K). Subsequently, government bonds are introduced as an additional form in which assets can be held. The rate of return on assets is denoted by r and is assumed to be paid out once per period. Expectations on r for future periods are assumed to be static at the current value. A member of the ith generation holds the amount of assets A_i^y while young and the amount A_i^o while old. The asset holding while old constitutes the provision of a bequest, which is assumed to go to the immediate descendant, a member of generation $i + 1$. Since the focus of the analysis concerns shifts in tax liabilities and government debt for a given level of government expenditure, it is assumed for convenience that the government neither demands commodities nor provides public services. In this section, it is also assumed that the amounts of government debt and taxes are zero. Using the letter c to denote consumption, and assuming that consumption and receipt of interest income both occur at the start of the period, the budget equation for a member of generation 1, who is currently old, is

$$A_1^y + A_0^o = c_1^o + (1 - r)A_1^o. \tag{1}$$

The total resources available are the assets held while young, A_1^y, plus the bequest from the previous generation, A_0^o. The total expenditure is consumption while old, c_1^o, plus the bequest provision, A_1^o, which goes to a member of generation 2, less interest earnings at rate r on this asset holding.

The budget equation for members of generation 2 (and, more g
for members of any generation $i \geq 2$) is, assuming that wage
occur at the start of the young period,

$$w = c_2^y + (1 - r)A_2^y,$$

and, for the old period,

$$A_2^y + A_2^o = c_2^o + (1 - r)A_2^o.$$

A portion of the lifetime resources of a member of generation i goes to a bequest provision, A_i^o, which I assume is motivated by a concern for a member of generation $i + 1$. This concern could be modeled by introducing either the (anticipated) consumption levels or attainable utility of a member of generation $i + 1$ into the utility function for a member of the ith generation. For the purpose of the present analysis, the crucial condition is that this utility depend on the endowment of a member of generation $i + 1$ rather than, per se, on the gross bequest, A_i^o. (The distinction between the gross bequest and the net bequest, which determines the endowment of $i + 1$, will be discussed below.) So long as a member of generation i can transfer resources to a member of generation $i + 1$ only through the transfer of unrestricted purchasing power (which rules out the "merit good" case discussed in n. 8 below), the two types of models of interdependent preferences—concern with consumption levels and concern with attainable utility—will be equivalent in the sense of indirectly implying a concern for the endowment of a member of generation $i + 1$.

For present purposes, it is convenient to assume that the utility of a member of generation i depends solely on own two-period consumption, c_i^y and c_i^o, and on the attainable utility of his immediate descendant, U_{i+1}^*. The asterisk denotes the maximum value of utility, conditional on given values of endowment and prices. Hence, the utility function for a member of the ith generation has the form,[4]

$$U_i = U_i(c_i^y, c_i^o, U_{i+1}^*). \tag{4}$$

Subsequently, I consider the implications of entering the attainable utility of a member of the previous generation, U_{i-1}^*, as an additional argument of the U_i function.

Each member of generation 1 determines his allocation of resources to maximize U_1, subject to equations (1)-(4) and to the inequality conditions, $(c_i^y, c_i^o, A_i^o) \geq 0$ for all i. The key restriction here is that the bequest to the member of the next generation cannot be negative.[5] The choice of bequest, subject to this restriction, takes into account the effect of A_1^o on generation 2's resources, the impact of U_2^* on U_1, and the chain dependence of U_2 on U_3^*, of U_3 on U_4^*, etc. The solution to this problem will take the general form

$$\begin{aligned} c_1^o &= c_1^o(A_1^y + A_0^o, w, r), \\ A_1^o &= \frac{1}{1-r}(A_1^y + A_0^o - c_1^o) = A_1^o(A_1^y + A_0^o, w, r). \end{aligned} \tag{5}$$

Similarly, for members of generation 2 (and, more generally, for members of any generation $i \geq 2$), the solution would take the form,

$$\begin{aligned} c_2^y &= c_2^y(A_1^o, w, r), \\ A_2^y &= \frac{1}{1-r}(w - c_2^y) = A_2^y(A_1^o, w, r), \\ c_2^o &= c_2^o(A_2^y + A_1^o, w, r), \\ A_2^o &= \frac{1}{1-r}(A_2^y + A_1^o - c_2^o) = A_2^o(A_2^y + A_1^o, w, r). \end{aligned} \tag{6}$$

---

[4] A member of generation i is assumed to be concerned with own consumption and with the attainable indifference surface of his descendant. Further, it is supposed that a member of generation i can attach a metric to generation $i + 1$'s indifference surface which makes it comparable to c_i^y and c_i^o in terms of generating U_i in the form of eq. (4). The nature of this sort of utility function is discussed in the general context of interdependent preferences in Becker (1974, sec. 3.A).

[5] I have not imposed the condition, $A_i^y \geq 0$, so that young individuals are allowed to issue interest-bearing debt on themselves. If issued, these debts are assumed to be perfect substitutes for equity capital. These debts correspond to the consumption loans which have been discussed by Samuelson (1958).

The model can be closed, as in Diamond (1965, pp. 1130–35), by specifying a constant-returns-to-scale production function that depends on the amounts of capital and labor input, and by equating the marginal products of capital and labor to r and w, respectively. The value of r for the current period would then be determined in order to equate the supply of assets to the demand—that is,

$$K(r, w) = A_1^o + A_2^y, \tag{7}$$

where $K(r, w)$ is such as to equate the marginal product of capital to r. The current demand for assets, $A_1^o + A_2^y$, depends, from equations (5) and (6), on r, w, and the previous period's value of K, which is equal to $A_1^y + A_0^o$. Since the number of people in each generation is assumed to equal a fixed number N, it is not necessary to enter this number explicitly into the aggregate asset demand in equation (7). Similarly, N is omitted from the aggregate formulations below. Since N is constant and technical change is not considered, the current and previous periods' values of K would be equal in a steady state.

With the marginal product of labor equated to w and with constant returns to scale, output is given by

$$y = rK + w. \tag{8}$$

Equations (2), (3), (7), and (8) imply a commodity mark condition,

$$c_1^o + c_2^y + \Delta K = y,$$

where ΔK denotes the change in capital stock from the previ current period. The value of ΔK would be zero in a steady stat present analysis is not restricted to steady-state situations.

B. Government Debt

Suppose now that the government issues an amount of debt, B, which can be thought of as taking the form of one-period, real-valued bonds. These bonds pay the specified amount of real interest, rB, in the current period and the specified real principal, B, in the next period.[6] It is supposed that asset holders regard equity and government bonds as perfect substitutes. It can be assumed, for simplicity, that the government bond issue takes the form of a helicopter drop to currently old (generation 1) households. Equivalently, it could be assumed that the bonds were sold on a competitive capital market, with the proceeds from this sale used to effect a lump-sum transfer payment to generation 1 households.

---

[6] The amount of bond issue would be limited by the government's collateral, in the sense of its taxing capacity to finance the interest and principal payments (see n. 12 below).

Allowing some portion of the proceeds to go to generation 2 households would not alter any of the basic conclusions.

The future interest payments on the government debt must be financed in some manner. Further, the principal may eventually be paid off—that is, the government may not reissue the bonds when they come due in the next period. I assume, provisionally, that the current period's interest payments are financed by a lump-sum tax levy on generation 2 households (while young), and that the principal is paid off at the beginning of the next period by an additional lump-sum tax levy on generation 2 households (while old). In this setup there is no direct effect of the government debt issue and its financing on generation 3 and later generations. I examine, subsequently, the implications of imposing some part of the taxes on generations of the more distant future.

The generation 1 budget constraint is now

$$A_1^y + A_0^o + B = c_1^o + (1 - r)A_1^o, \tag{10}$$

where B represents the lump-sum transfer payment, which is assumed to occur at the beginning of the period. For generation 2, the current budget constraint is now

$$w = c_2^y + (1 - r)A_2^y + rB, \tag{11}$$

where rB represents the tax levy for the government interest payments. The next period's budget constraint for generation 2 is now

$$A_2^y + A_1^o = c_2^o + (1 - r)A_2^o + B,$$

where B represents the tax levy for repayment of principal. The two constraints on generation 2 can be combined into a single two-period budget equation,

$$w + (1 - r)A_1^o - B = c_2^y + (1 - r)c_2^o + (1 - r)^2A_2^o. \tag{12}$$

The form of equation (12) implies that the utility attainable by a member of generation 2 can be written in the indirect form,

$$U_2^* = f_2^*[(1 - r)A_1^o - B, w, r], \tag{13}$$

that is, the "net bequest," $(1 - r)A_1^o - B$, determines the "endowment" for members of generation 2.

From equation (10), it is also clear that c_1^o varies inversely with $(1 - r)A_1^o - B$ for a given value of $A_1^y + A_0^o$. Hence, given the predetermined value of c_1^y, and using equations (4), (10), and (13), U_1 can be written in the form,

$$U_1 = U_1(c_1^y, c_1^o, U_2^*) = f_1[(1 - r)A_1^o - B; c_1^y, A_1^y + A_0^o, w, r].$$

For given values of c_1^y, $A_1^y + A_0^o$, w, and r, the choice problem for members of generation 1 amounts to the optimal selection of the net bequest,

$(1 - r)A_1^o - B$, subject to the constraint that the gross bequest, A_1^o, be nonnegative. In particular, if the solution to this problem is associated with a value of A_1^o in the interior—that is, if the constraint, $A_1^o \geq 0$, is not binding—any marginal change in B would be met solely by a change in A_1^o that maintains the value of the net bequest, $(1 - r)A_1^o - B$. This response in A_1^o will keep unchanged the values of c_1^o, c_2^y, c_2^o, and A_2^o. Hence, the utility levels attained by members of generations 1, 2, etc., will be unaffected by the shift in B.

In terms of the effect on r, the current asset market clearing condition of equation (7) would now be modified to

$$K(r, w) + B = A_1^o + A_2^y. \tag{14}$$

The increase in B implies a one-to-one increase in the asset supply on the left-hand side of equation (14). However, A_1^o rises by $1/(1 - r)$ times the change in B in order to maintain the size of the net bequest, $(1 - r)A_1^o - B$. Further, with c_2^y fixed, the increase in rB (taxes) in equation (11) implies that A_2^y falls by $r/(1 - r)$ times the change in B. On net, total asset demand on the right-hand side of equation (14) rises one-to-one with B, so that no change in r is required to clear the asset market. Equivalently, the commodity market clearing condition, as expressed in equation
continues to hold at the initial value of r because the bond is
impact on aggregate demand.

Essentially, a positive value of B, financed by a tax levy
generation, enables a member of the old generation to "go ou
by leaving a debt for his descendant. However, if, prior to the
bond issue, a member of the old generation had already selecte
bequest, it is clear that this individual already had the option
resources from his descendant to himself, but he had determined that
shifting, at the margin, was nonoptimal. Since the change in B does not alter the relevant opportunity set in this sense, it follows that—through the appropriate adjustment of the bequest—the values of current and future consumption and attained utility will be unaffected. On the other hand, if a member of generation 1 were initially at a corner where $A_1^o = 0$—in particular, if $A_1^o < 0$ would have been chosen had it been permissible—then an increase in B creates a relevant new opportunity. In this situation a generation 1 household would react by increasing c_1^o along with B, as long as the corner solution for A_1^o still applied. The upward shift in B would then correspond to an excess of earning-asset supply over demand (even after taking account of a shift in A_2^y), which would tend to raise the value of r. This increase in r would induce a drop in capital formation, which constitutes the real effect of government debt issue which has been described by Modigliani (1961). However, the main point is that the existence of this government debt effect hinges on a non-

operative bequest motive—that is, on households being at the corner where the amount of bequest is zero.[7]

It should be stressed that the crucial consideration for the above result is an operative intergenerational transfer, rather than an operative bequest motive per se. For example, the transfer could take the form of parental expenditure on children's education, etc., during the overlapping tenure of parent and child.[8] Further, the transfer could be occurring in the direction opposite to that specified above. In particular, U_1^* could be entered as an argument of the U_2 function, and the possibility of gifts from the young to the old generation could be introduced. In that case the same conclusions on the effect of a change in the government debt would be reached if a "gift motive" were operative.[9] The mechanism through which changes in B were offset would then be an alteration in the amount of gifts from young to old, rather than an alteration of the amount of bequests from old to young.

The results will now be extended to a situation where the taxes which finance the government debt affect some generations which are not currently alive. The extension will be made explicitly only to generation 3, since the extension to generations further advanced in the future is straightforward.

Suppose now that the current period's interest payments are financed by a lump-sum tax levy on (young) generation 2, the next period's interest payments (on the reissued bonds) are financed by a lump-sum tax levy on (young) generation 3, and the principal is paid off by a lump-sum tax levy on (old) generation 3.[10] The generalization of the earlier results to this situation can be demonstrated by working backward from generation 3. By analogy to equation (13), the attainable utility of generation 3 can be written in the indirect form,

---

[7] When households are not identical, the aggregate effect of government debt issue will depend on the fraction of households at a corner. As long as some households are in this situation, a shift in B will have some upward effect on r in this model. However, this effect would be "small" if the fraction of households at a corner were small. The role of a bequest motive in eliminating the perceived net-wealth effect of government debt has also been discussed by Miller and Upton (1974, pp. 176–79).

[8] The previous results on the effect of B might not hold if parents were concerned with specific consumption components of their children ("merit goods"), rather than with their children's attainable utility. Formally, U_i in eq. (4) could depend on (components of) c_{i+1}^y or c_{i+1}^o, rather than on U_{i+1}^*. If generation i can tie its aid to generation $i + 1$ to a specific type of expenditure (as could be the case for education), the previous results would not hold if this tied aid were an effective constraint—in the sense of forcing the next generation to "purchase" more of the item than it otherwise would—and if the parents were not making any other transfers which were equivalent to the transfer of general purchasing power. Becker (1974, sec. 3.C) presents a detailed discussion of the merit goods case in an analogous context.

[9] A model which allows for a reciprocal dependence between U_i and U_{i+1} is formally similar to the model discussed by Becker (1974, sec. 3.A) in the context of transfer payments among members of a family.

[10] I do not deal here with the possibility of net government debt issue during the old-age tenure of generation 2. No new considerations would arise here (see however, n. 12 below).

$$U_3^* = f_3^*[(1 - r)A_2^o - B, w, r],$$

where $(1 - r)A_2^o - B$ now determines the endowment for members of generation 3. Since generation 2 no longer pays off the government debt principal, its budget equation is modified from the form of equation (12) to

$$w + (1 - r)A_1^o - B = c_2^y + (1 - r)c_2^o + (1 - r)[(1 - r)A_2^o - B].$$

For given values of w, r, and the net bequest from generation 1, $(1 - r)A_1^o - B$, generation 2 would select an optimal value of the net bequest to generation 3, $(1 - r)A_2^o - B$. This net bequest would be invariant with B as long as the solution for A_2^o were interior. Assuming that this solution is interior, the attainable utility of generation 2 can be written in the indirect form,

$$U_2^* = f_2^*[(1 - r)A_1^o - B, w, r],$$

which coincides in form with equation (13). The situation has therefore been reduced to the previous case in which marginal changes in B led solely to changes in A_1^o which kept $(1 - r)A_1^o - B$ constant without affecting any values of consumption or attained utility.

The three-generation results generalize to the case in which taxes are levied on m generations, with the mth generation paying off the principal By starting with generation m and progressing backward, it can b for all $2 \leq i \leq m - 1$ that, if A_i^o is interior, U_i^* can be wri indirect form as a function of $(1 - r)A_{i-1}^o - B$. As long as all i choices are interior[11] (as anticipated by current generations), imply fully compensating shifts in bequests, so as to leave unc values of consumption and attained utility.[12]

---

[11] Intuitively, if this condition is violated for some generations, the impa violations on current behavior should be less important the further in the futur violating generations. I make no claim to having proved this conjecture.

[12] This line of proof does not apply as $m \rightarrow \infty$. The main issue seems to be whether the assumption that the principal is eventually paid off is crucial. If the amount of outstanding government debt were constant, the impact of the principal on current decisions would become negligible for large m as long as $r > 0$. However, a difficulty arises here when B is allowed to grow over time. Suppose that the growth of B were limited to the growth of the government's collateral in the sense of its taxing capacity, which depends in turn on the growth of real income. Suppose that the growth rate of real income is equal to n, which can be viewed as the combined effects of population growth and technical progress, which are now allowed to be positive. In that case the present value of the principal would have to become negligible as $m \rightarrow \infty$ if $n < r$. The situation in which $n > r$ applies is inefficient in that it is associated with a capital stock in excess of the golden rule level (see, e.g., Diamond 1965, p. 1129). It is possible in Diamond's model (p. 1135) that the competitive equilibrium can be in this inefficient region. However, this situation is not possible in growth models where individuals are infinite lived and utility is discounted (see, e.g., Koopmans 1965). As long as intergenerational transfers are operative, the overlapping-generations model would seem to be equivalent to the infinite-life model in this respect — that is, the possibility of inefficiency in Diamond's model seems to hinge on finite lives with inoperative intergenerational transfers. Hence, when these transfers are operative, $n < r$ would be guaranteed, and the possibility of perpetual government finance by new debt issue could then be ruled out.

The results in this section have demonstrated that changes in government debt would not induce any alteration in consumption plans even in a model where (1) the present generations have finite lives, (2) the present generations may, in some sense, give lesser weight to the consumption or utility of future generations than they give to own consumption, and (3) the present generation may give no direct weight at all to the consumption or utility of generations beyond their immediate descendants (who are also finite-lived).

A sufficient condition for changes in government debt to have no impact on consumption plans and, hence, no effect on aggregate demand and interest rates is that the solution for the current generations' inheritances be interior, and that the solutions for future generations' inheritances (as perceived by current generations) also be interior. More generally, the result will hold as long as current generations are connected to all future generations by a chain of operative intergenerational transfers, either in the direction from old to young or in the direction from young to old.

The derivation of conditions under which the solution for intergenerational transfer would be interior appears to be a difficult problem and would seem to require some specialization of the form of the utility functions in order to make any headway. However, it seems clear that bequests are more likely to be positive the smaller the growth rate of w (assuming that w is now viewed as variable across generations), the higher the interest rate, the higher the relative weight of U_{t+1}^* in the U_t function, and the larger the value of B.[13] The reverse conditions favor a gift from young to old.[14]

C. Social Security Payments and Other Imposed Intergenerational Transfers

The above results on government debt also apply to social security payments.[15] Suppose that a scheme is instituted which immediately begins payments to the current old generation (generation 1) of amount S, financed by a lump-sum tax levy of amount S on the current young generation (generation 2). Generation 2 expects to receive a transfer of amount S while old, financed by a lump-sum tax levy on (young) generation 3, etc. It is assumed here that an individual's payment received while old is independent of his own contribution to the scheme while young, and that neither the old receipt nor the young payment depends on the amount of work, income, etc. Assuming interior bequests (which would be guaranteed by a sufficiently high value of S), a change in S would induce the current old generation (generation 1) to maintain its choice of c_1^o and, correspondingly, to raise A_1^o by $1/(1 - r)$ times the change in S. This increased inheritance would just offset the increased tax liability imposed on (young) generation 2. With its consumption unchanged, generation 2 would use its own higher social security receipt to raise its bequest to generation 3, A_2^o, by $1/(1 - r)$ times the change in S. As in the case of changes in government debt, if the solutions for bequest are interior, the impact of a marginal change in S would be solely on the size of bequests and not at all on the pattern of consumption.[16] The same results would follow in the case of operative intergenerational transfers from young to old, with a marginal increase in S implying a corresponding reduction in the size of gifts from young to old.

---

[13] In a more general context B should be viewed as outstanding public debt less the value of physical capital held by the government.

[14] There is an alternative argument, which Gary Becker refers to as the "enforcement theory of giving," which suggests that bequest motives would typically be operative. Suppose that, instead of receiving utility from the perceived utility of his child, a parent is concerned with own consumption and with the amount of attention, etc., shown by his child during their overlapping tenure. Suppose, further, that the child has some information on the size of his parents' estate and that—acting as a good optimal controller—he regulates the amount of attention as a function of the estate size. In this situation the estate would surely be positive if parents place a high value on getting at least a small amount of attention, and if the child provides no attention when the estate is zero. However, although a positive estate could be guaranteed in this fashion, it seems that the previous conclusions about the marginal effect of B on consumption plans would not hold in this model. The nature of the interactions between parents and children would have to be analyzed more fully for this case.

[15] The view of social security as analogous to government debt has also been taken by Miller and Upton (1974, pp. 182–84).

The results for social security payments would apply also to other programs which amount to imposed intergenerational transfer schemes. In particular, public support of education involves a forced transfer [illegible] resources from old to young. In the main, this sort of impo[illegible] would be offset by adjustments in the opposite direction of [illegible] transfers.[17]

D. Inheritance Taxes

Suppose now that inheritances (or gifts) are taxed at a proportional [illegible] τ. In particular, the bequest from a member of generation i, A_i^o, yield [illegible] net receipt to his descendant, a member of generation $i + 1$, of size $(1 - \tau)A_i^o$. Of course, the tax receipts must also go somewhere. Suppose that these receipts are transferred to members of generation $i + 1$ (while old) in accordance with a rule that is independent of the size of each individual's inheritance.

---

[16] As in the case of government debt issue, the formal proof depends on the assumption that the scheme is eventually liquidated (see n. 12 above). The consumption patterns would also not be affected by a social security scheme that involved the accumulation of a government "trust fund." Assuming that the fund were held in the form of earning assets, an increase in the fund would be equivalent to a negative government debt issue. Real effects of a social security system would arise if the payments were contingent on the work behavior of the old generation. In that case there would be allocative effects produced by the disincentive to work in later years.

[17] On a theoretical level, government education programs will involve real effects to the extent that (1) there is an efficiency difference between public and private production of education, (2) public expenditure on education is pressed sufficiently far so that a reduction of discretionary transfers cannot occur on a one-for-one basis, and (3) there are distributional effects involving relative educational expenditures and tax liabilities across families. As an empirical matter, Peltzman (1973) has shown that public subsidies for higher education are offset to an extent of about 75 percent by reductions in private expenditures for higher education. However, Peltzman's 75 percent figure does not coincide with the desired estimate of the effect on discretionary transfers, since other components of discretionary transfers may also be affected and (on the other side) since not all private expenditures for education constitute intergenerational transfers.

Since an individual's contribution to general tax revenue will typically be valued by him at less than an equal amount of own income, it is clear that an increase in τ will tend to lower the amount of intergenerational transfers. In particular, the higher the value of τ, the less likely that a bequest or gift motive will be operative. Suppose, however, that the value of τ is sufficiently low that all intergenerational transfers are operative, even if at reduced levels. In this case the previous results on the effect of a change in government debt remain valid.

Consider the situation in which the principal on the government debt is paid off by generation 2. Equation (10) continues to apply in the presence of inheritance taxes, but equation (12) must be modified to

$$w + (1 - r)(1 - \tau)A_1^o + (1 - r)\tau\overline{A_1^o} - B$$
$$= c_2^y + (1 - r)c_2^o + (1 - r)^2A_2^o,$$

where $\tau\overline{A_1^o}$ represents the transfer to a member of (old) generation 2 corresponding to his share of the receipts from the total taxes paid on the average generation 1 bequest, $\overline{A_1^o}$. In deciding on a plan for consumption and intergenerational transfers, an individual is assumed to treat $\tau\overline{A_1^o}$ as exogenous. Consider the conjecture that, when B rises, each member of generation 1 continues to respond by maintaining the value of c_1^o and, hence, by maintaining the value of the net pretax bequest, $(1 - r)A_1^o - B$. This response requires an increase in A_1^o by $1/(1 - r)$ times the increase in B. Each individual's net posttax bequest would fall in this case, but this fall would be offset, at least on average, by an increase in the transfers to generation 2 which are financed from the inheritance tax receipts, $\tau\overline{A_1^o}$. In this circumstance, the individual values of c_2^y, c_2^o, and A_2^o– and, hence, the attained value of U_2 --would remain fixed. Hence, by maintaining the net pretax bequest, each member of generation 1 achieves the same combination of c_1^o and U_2^* as before the shift in B. On the other hand, if an individual member of generation 1 decided to increase his net pretax bequest, while all other members held their net pretax bequests fixed, it would turn out for this individual that U_2^* would increase, while c_1^o would decrease. The terms on which an individual can exchange c_1^o for U_2^* depend on τ and r, and these terms have not been altered by the change in B. Further, when the transfer to generation 2 of size $\tau\overline{A_1^o}$ is included, there is also no change in an individual's overall wealth position. Therefore, the pattern which maintains the net pretax bequest—and thereby involves no shift in c_1^o or U_2^*—must be the optimal pattern for an individual. It follows that constancy of the net pretax bequest for all members of generation 1 is the equilibrium solution.[18] In this case, a marginal shift in B again has no effect on consumption patterns.

The basic conclusion here is that the existence of taxes on intergenerational transfers makes less likely an interior solution for these transfers, but if these transfers are operative, even if at reduced levels, the marginal effect of B on consumption plans—and, hence, on r—remains nil.

E. Bond Issue and Tax-Collection Costs

Suppose now that the issue of government debt and the collection of taxes to finance this debt involve transaction costs. In particular, in the case where the principal is paid off by generation 2, suppose that a net issue of B to generation 1 is now associated with a tax levy of $(1 + \gamma)rB$ on (young) generation 2 and a levy of $(1 + \gamma)B$ on (old) generation 2. That is, γ amounts to a proportional transaction cost associated with government debt issue and tax collection.[19] For simplicity, suppose now that the inheritance tax rate is zero. Equation (10) again remains valid, but equation (12) is now modified to

$$w + (1 - r)A_1^o - (1 + \gamma)B = c_2^y + (1 - r)c_2^o + (1 - r)^2 A_2^o$$

Consider, again, the conjecture that, when B rises, c_1^o an $(1 - r)A_1^o - B$ remain fixed. From equation (15), $\gamma > 0$ negative-wealth effect on generation 2, so that U_2^* would fall effect would be anticipated by generation 1, it can be supp normal case that A_1^o would actually rise by somewhat n $1/(1 - r)$ times B, so that c_1^o would fall. In general, $\gamma > 0$ implie increase in B amounts to an overall negative-wealth effect, which wou typically involve reductions in both c_1^o and U_2^*. This effect can be seen by combining equations (10) and (15) into the single two-generation budget equation,

---

[18] The equilibrium satisfies two properties: (1) each individual chooses his bequest optimally, subject to a given choice of bequests by all other individuals, and (2) all individuals choose the same value for their bequests. It can also be shown that the solution that maintains the net pretax bequest for all individuals is the unique equilibrium. Finally, it can be noted that the solution involves the assumption that each individual perceives the shift in the transfer term, $\tau\bar{A}_1^o$, associated with the average response of bequests to the change in B. Alternatively, if individuals treated $\tau\bar{A}_1^o$ as fixed, they would view an increase in B as, effectively, a negative change in wealth. The typical response would be a reduction in c_1^o, which would be associated with an increase in A_1^o by more than $1/(1 - r)$ times the change in B. In the aggregate, there would be an increase in desired saving, $A_2^o + A_1^y$, which would lead to a reduction in r and to an increase in capital formation. In particular, if the shift in transfers associated with inheritance tax revenues, $\tau\bar{A}_1^o$, is not perceived, the effects would be opposite to the standard case in which perceived net wealth rises with B.

[19] If the initial debt issue is associated with a decrease in other taxes, rather than an increase in transfers, there could be an offsetting reduction in transaction costs. The parameter γ, which is assumed to be positive, must be interpreted in this net sense.

$$A_1^y + A_0^o - \gamma B + w = c_1^o + c_2^y + (1 - r)c_2^o + (1 - r)^2 A_2^o. \quad (16)$$

The decline in total resources on the left-hand side of equation (16) produced by an increase in B would typically be reflected in declines in all terms on the right-hand side—c_1^o, c_2^y, c_2^o, and A_2^o.

In this circumstance the effect on r of a shift in B would be unclear. The commodity market clearing condition of equation (9) would now be modified to include the resources devoted to bond and tax transactions. The revised market clearing condition would be

$$c_1^o + c_2^y + \Delta K + \gamma r B = y.$$

The effect of B on current r will depend on whether, for a given value of r, the sum, $c_1^o + c_2^y$, falls by more or less than the increase in $\gamma r B$. This relationship seems to be ambiguous.[20]

II. Imperfect Capital Markets

This part of the paper analyzes the implications of divergences among individual discount rates. This source of a net-wealth effect for government bonds has been stressed by Mundell (1971), who argues that, because of high discount rates for some individuals, the taxes which finance the government debt will not be fully capitalized – hence, an issue of government bonds will involve a net-wealth effect. To analyze this effect, it is necessary to construct a somewhat different model. Suppose that there are now two types of individuals – those who have a low discount rate, r_l, and those who have a high discount rate, r_h. It can be supposed that the high-discount-rate individuals have relatively "bad collateral," so that loans to these individuals involve high transaction costs, which are reflected in high (net-of-default-risk) borrowing rates.[21] In particular, suppose that the two discount rates are related according to

$$r_h = (1 + \lambda)r_l,$$

where $\lambda > 0$ represents the proportional transaction costs involved in the loan process.[22] I suppose in this part of the paper that both types of individuals are infinite-lived, since the effect of finite lives has already been examined above.

---

[20] From eq. (16), the negative wealth effect is γB, which is the present value of the flow, $\gamma r B$. The sum, $c_1^o + c_2^y$, will fall by as much as $\gamma r B$ if the total "propensity to consume" associated with the negative "income" flow, $\gamma r B$, is equal to one.

[21] In this respect see Barro (1974).

[22] I am assuming that the r_h individuals are actually borrowing, so that r_h represents both their borrowing rate and their marginal discount rate. Alternatively, r_h could be viewed as a marginal discount rate which could be somewhere between the borrowing and lending rates, as in Hirshleifer (1958).

It is convenient to suppose that government debt now takes the form of a perpetuity that carries a real interest payment of i per year. Suppose that the government issues an additional bond of this type. This bond would be purchased by a low-discount-rate individual and would be evaluated as $B = i/r_l$.[23] Suppose then that the government uses the lump-sum proceeds from this sale, B, to effect a lump-sum transfer (or lump-sum tax reduction) to individuals, and suppose that a fraction α of this transfer goes to r_l discount rate individuals and a fraction $(1 - \alpha)$ to r_h discount rate individuals. Finally, the taxes for financing the government interest payments are $(1 + \gamma)i$, where γ represents, as in section I*E*, the proportional transaction costs associated with government bond sale and tax collection. Suppose that these taxes are distributed across discount rates in the same manner as the lump-sum proceeds[24]—that is, a fraction α to r_l individuals and a fraction $(1 - \alpha)$ to r_h individuals.

Consider, in turn, the wealth effects for the r_l and r_h groups. The bond sale itself involves no wealth effect for the r_l group. The lump-sum transfer to r_l individuals is $\alpha B = \alpha i/r_l$, while the present value of the r_l share of tax liabilities, discounted at rate r_l, is $(1 + \gamma)\alpha i/r_l$. Clearly, if $\gamma > 0$, the net-wealth effect for r_l individuals is negative, as it was in the case discussed in section I*E*, where all discount rates were equal.

For the r_h group, the lump-sum proceeds are $(1 - \alpha)B =$
while the present value of the tax liability, discounted at
$(1 + \gamma)(1 - \alpha)i/r_h$. Using $r_h = (1 + \lambda)r_l$, the net-wealth effe
be expressed as

$$\frac{(1 - \alpha)i}{r_l}\left(1 - \frac{1 + \gamma}{1 + \lambda}\right) = \frac{(1 - \alpha)i}{r_l(1 + \lambda)}(\lambda - \gamma),$$

which is positive if $\lambda > \gamma$. That is, the net-wealth effect for the r_h group is positive if γ, which measures the government transaction costs for bond issue and tax collection, is smaller than λ, which measures the private transaction costs implicit in the existing pattern of (net-of-default-risk) discount rates. To the extent, $1 - \alpha$, that the transfer payment and tax liability involve the r_h group, the government bond issue amounts to effecting a loan from the low-discount-rate to the high-discount-rate individuals. On the other hand, this sort of transfer could already have been accomplished privately, except that the transaction costs, as measured by λ, made this transfer marginally unprofitable. Hence, the government-induced transfer implied by its bond issue can raise net wealth only if the government is more efficient than the private capital market in carrying out this sort of lending and borrowing operation.

---

[23] This analysis abstracts from any "liquidity yield" of bonds (see part III, below).

[24] If the fractions for transfer and tax liability vary, then the wealth effects on the two discount-rate groups are likely to be in opposite directions. The net effect on current consumption demand would depend, in part, on relative propensities to consume, which are not obvious. In any event, this case would amount to the effect of income distribution on consumption demand, rather than the effect of government bond issue *per se* on net wealth and consumption demand.

Some additional observations can be made concerning this result. First, if the government is really more efficient than the private market in the lending process (presumably because the benefits of economies of scale [in information?] and the ability to coerce outweigh the problems of government incentive and control), it may be able to exploit this efficiency better by a direct-loan program, rather than by the sort of bond issue described above. In my simple model, a fraction α of transfers and tax liabilities involved the r_l group, and this process entailed a dead-weight loss to the extent that $\gamma > 0$. A program which limited the loan recipients to high-discount-rate individuals would be more efficient in this respect. However, the information requirements for this sort of program may be much greater than those for a program which does not attempt to discriminate—in the transfer and tax liability aspects—among discount rates. The crucial point which can make the bond issue work as a loan program is that the purchasers of the bonds automatically discriminate among themselves as to their discount rates.

Second, the government may be more efficient than the private market only over a certain range of B. In particular, there may be a sufficiently large value of B such that, at the margin, the net-wealth effect of government debt is zero. If the public choice process leads to this value of B (as it should on efficiency grounds), then, at the margin, the net-wealth effect of government bonds would be zero, despite the continued existence of "imperfect private capital markets."[25]

III. A Government Monopoly in Liquidity Services

Suppose now that government debt provides a form of "liquidity service" to the holder, in addition to the direct interest payments. Suppose that, at the margin, these services are valued at the amount L per bond per year. Hence, in the context where all individuals have the same discount rate, r, an additional perpetual government bond would be evaluated as

$$B = (1 + L)/r.$$

The taxes for financing the government debt can be thought of as the interest costs, 1, plus any costs involved with the process of creating liquidity services (which could involve the γ-type costs discussed above). Suppose that c denotes the marginal costs per bond per year associated with the production of liquidity services. Hence, at the margin, the wealth effect of a change in government debt will be

---

[25] Of course, government debt issue would be "productive" in a total sense even in the case where the marginal net wealth effect was nil. However, it is this marginal effect which enters into analyses of (marginal) fiscal and monetary policies.

$$\frac{1}{r}(i + L) - \frac{1}{r}(i + c) = \frac{1}{r}(L - c).$$

If the public choice process is such as to motivate the government to act like a competitive producer of liquidity services (as it should on efficiency grounds), then $L = c$ and the marginal-wealth effect of government debt would be nil. On the other hand, if the government operates monopolistically, so that $L > c$, then the marginal-wealth effect of government debt would be positive.[26] However, it is also possible that the government overextends its production of liquidity services, so that $L < c$ and the marginal-wealth effect of government debt would be negative. This last case corresponds implicitly to the one discussed above in section I*E*, where $L = 0$ and $c > 0$ were assumed.

Of course, liquidity services can also be provided by private producers. If the types of services rendered by private and public debt instruments are close substitutes, and if the private market is competitive, then governmental monopoly power can arise here only to the extent that, at the margin, the government is more efficient than the private mark producer of liquidity services. Even if the government is a mo producer over a certain range, a sufficient expansion of g "output" would eliminate this efficiency differential at the m production of liquidity services is, at least eventually, subject to marginal costs. As in the case of an imperfect private capital discussed above, the net-wealth effect of government debt dep relative efficiency at the margin of government versus private produ

IV. Risk and Asset Substitutability

The previous sections have dealt with the net-wealth effect of government debt. I have not discussed explicitly in these sections the risk characteristics of government bonds, tax liabilities, and the other types of available assets and liabilities. Tobin (1971, p. 2) has argued: "The calculus of total wealth is less important than the change in the composition of private balance sheets that the government engineers by borrowing from the public—forcing on taxpayers a long-term debt of some uncertainty while providing bond-holders highly liquid and safe assets. Since no one else can perform the same intermediation, the government's debt issues probably do, within limits, augment private wealth. Another way to make the point is to observe that future tax liabilities are likely to be capitalized at a higher discount rate than claims against the government." I have already considered, above, arguments for effectively discounting tax liabilities at a higher rate because of finite lives, imperfect private capital markets, and a government monopoly in the production of liquidity services, and these arguments need not be repeated here. In this part of the paper, I will consider briefly some implications of the risk characteristics of government bonds and of the future tax liabilities associated with the finance of these bonds.

---

[26] Of course, this observation would also apply to government money, which yields a zero rate of explicit interest. The usual real balance effect for outside money assumes that the marginal cost to the government of maintaining real balances is zero, and that the government acts like a monopolist in determining its supply of real balances.

Suppose, first, that there were no uncertainty about the relative burden of the (lump-sum) tax liabilities that finance the government debt. In this situation the uncertainty in an individual's real tax burden associated with government interest payments would reflect solely the variability over time in the real-interest payments themselves. In terms of present values, the variability in the tax liabilities would reflect the variability in prices and interest rates—that is, the same factors which lead to variability in real bond values. In particular, holdings of government debt—amounting to a claim to a certain fraction of total government interest payments—would be the perfect hedge against variations in tax liabilities.[27] In this context a simultaneous increase in government interest payments (i.e., government bonds) and in the tax liabilities for financing these payments would not involve any net shift in the risk composition of private balance sheets.[28]

Suppose now that the tax liabilities are subject to an additional variability concerning the relative burden across individuals. Suppose, first, that the variation in relative taxes is purely random, in the sense of being unrelated to variations in relative income, etc. In that case, it is clear that an individual's tax liability associated with government interest payments would be subject to a source of variability above that of the total interest payments. In particular, the fractional holdings of government bonds which corresponds to the expected fraction of tax liabilities would no longer provide a perfect hedge against variations in the tax liabilities. Of course, it would be possible for individuals to utilize private insurance markets to reduce the risks associated with variations in relative tax liability. However, to the extent that insurance arrangements entail transaction costs, the risk associated with relative liability would not generally be fully eliminated. In this case an increase in government bonds would produce a net increase in the risk contained in household balance sheets—that is, there would be a decline in effective household wealth. The typical household reaction would be twofold: first, an increase in desired total saving, and, second, a shift in portfolio composition away from more risky assets, such as equity capital, and toward less risky assets. The impact on the equity rate of return, and, hence, on capital formation would depend on which of these two responses was the dominant force.

---

[27] I am ignoring here effects which relate to the maturity structure of the government debt. In order to provide a perfect hedge, an individual's holding of debt by maturity would have to correspond to the overall maturity distribution.

[28] There could be an effect on individuals who do not hold any government bonds (or assets subject to similar risks).

The above discussion would be altered to the extent that variations in tax liability reflect variations in income. In this context the variation in relative tax liabilities can serve to reduce the net variability in disposable income—that is, the income tax works, in part, like a public program of income insurance. If the income-offsetting feature of taxes were the dominant element in relative tax variability, then a shift in government bonds could lead to a reduction in the overall risk contained in household balance sheets. In that case the effects on desired total saving and on portfolio composition would be opposite to those described above. However, it should also be noted that the public program of income insurance which is implied by an income tax system will also involve transaction costs. There are costs associated with administration and with individual reporting effort, as well as "moral hazard" costs associated with incentives for earning income. A full analysis of the wealth effect of government bonds under different tax systems would have to involve a comparison of these types of public transaction costs against the tran costs associated with the pooling of income risks under private i arrangements.

One final observation can be made here. The argument in t literature for a net-wealth effect of government bonds—for exam given by Modigliani (1961)—involved a neglect of the tax l associated with the financing of the debt. Similarly, Tobin's argum effects based on the risk composition of household balance sheets seems neglect the tax liabilities as an element of these balance sheets. It seems clear that, either in the sense of effects on perceived total wealth, or in the sense of the risk composition of household portfolios, the impact of changes in government debt cannot be satisfactorily analyzed without an explicit treatment of the associated tax liabilities. Once the variability in relative tax liability is considered, there seem to be no clear results concerning the effect of government debt issue on the overall risk contained in household balance sheets. The net effect hinges on the extent to which variations in relative tax liability reflect variations in relative income, and on the transaction costs for public programs of income insurance relative to those of private programs.

V. Summary and Conclusions

This paper has focused on the question of whether an increase in government debt constitutes an increase in perceived household wealth. The effect of finite lives was examined within the context of an overlapping-generations model of the economy. It was shown that households would act as though they were infinitely lived, and, hence, that there would be no marginal net-wealth effect of government bonds, so long as there existed an operative chain of intergenerational transfers which connected current to future generations. Net-wealth effects associated with imperfect private capital markets and with a government monopoly in the production of liquidity services were shown to depend on the assumption that the government was more efficient, at the margin, than the private market either in the loan process or in the production of liquidity services. Further, the introduction of government transaction costs for bond issue and tax collection implied that the net-wealth effect of government bonds could be negative. Finally, a consideration of the risk characteristics of government debt and of the tax liabilities associated with the financing of this debt suggested that an increase in government bonds could raise the overall risk contained in household balance sheets. However, this effect depends on the nature of the tax system and on the transaction costs associated with private insurance arrangements.

The basic conclusion is that there is no persuasive theoretical case for treating government debt, at the margin, as a net component of perceived household wealth. The argument for a negative wealth effect seems, a priori, to be as convincing as the argument for a positive effect. Hence, the common assertion (as in Patinkin 1962, chap. 12, p. 289) that the marginal net-wealth effect of government bonds is somewhere between zero and one and is most likely to lie at some positive intermediate value has no a priori foundation. If, in fact, the marginal net-wealth effect were negligible, the implications for monetary and fiscal analysis would be far-reaching. In particular, in the case where the marginal net-wealth effect of government bonds is close to zero, (1) the Metzler-type argument for nonneutrality of changes in the stock of outside money would not be valid, (2) a change in the stock of government debt would have no effect on capital formation, and, more generally, (3) fiscal effects involving changes in the relative amounts of tax and debt finance for a given amount of public expenditure would have no effect on aggregate demand, interest rates, and capital formation.[29]

References

Bailey, M. J. *National Income and the Price Level.* New York: McGraw-Hill, 1962.

Barro, R. J. "The Loan Market, Collateral, and Rates of Interest." Center for Math. Studies in Bus. and Econ., Univ. Chicago, Report 7401, January 1974.

---

[29] The usual fiscal analysis involves a shift in the flow of government debt rather than a one-time shift in the stock. The zero net-wealth effect applies also to the flow case if individuals perceive the implications of the current flow for the future time path of the stock of government debt.

Becker, G. S. "A Theory of Social Interactions." *J.P.E.* 82, no. 6 (November/December 1974): 1063–93.

Blinder, A. S., and Solow, R. M. "Does Fiscal Policy Matter?" *J. Public Econ.* 2 (November 1973): 319–37.

Diamond, P. A. "National Debt in a Neoclassical Growth Model." *A.E.R.* 60 (December 1965): 1126–50.

Hirshleifer, J. "On the Theory of Optimal Investment Decisions." *J.P.E.* 66, no. 4 (August 1958): 329–52.

Koopmans, T. C. "On the Concept of Optimal Economic Growth." In *The Econometric Approach to Development Planning.* Amsterdam: North-Holland, 1965.

Metzler, L. "Wealth, Saving, and the Rate of Interest." *J.P.E.* 59, no. 2 (April 1951): 93–116.

Miller, M. H., and Upton, C. W. *Macroeconomics: A Neoclassical Introduction.* Homewood, Ill.: Irwin, 1974.

Modigliani, F. "Long-Run Implications of Alternative Fiscal Policies and the Burden of the National Debt." *Econ. J.* 71 (December 1961): 730–55.

Mundell, R. "Money, Debt, and the Rate of Interest." In *Monetary Theory,* edited by R. Mundell. Pacific Palisades, Calif.: Goodyear, 1971.

Patinkin, D. *Money, Interest, and Prices.* 2d ed. New York: Harper & Row, 1964.

Peltzman, S. "The Effect of Public Subsidies-in-Kind on Private Expenditures: The Case of Higher Education." *J.P.E.* 81, no. 1 (January/February 1973): 1–27.

Samuelson, P. A. "An Exact Consumption-Loan Model of Interest with or without the Social Contrivance of Money." *J.P.E.* 66, no. 6 (December 1958): 467–82.

Thompson, E. A. "Debt Instruments in Macroeconomic and Cap
A.E.R. 67 (December 1967): 1196–1210.

Tobin, J. *Essays in Economics.* Vol. 1. *Macroeconomics.* Amsterdam:
1971.

Wage Comparisons—a Selectivity Bias

Reuben Gronau

The Hebrew University and National Bureau of Economic Research

The wage rate a person receives depends not only on the wage offered (a function of his market characteristics), but also on his job-search strategy. The higher his wage demands, the higher the wage he can expect, though the probability of finding an adequate job is lower. When comparing wages of different population groups, one must take their different search policies into account. Ignoring these differences results in a selectivity bias. This bias is particularly relevant for the secondary labor groups; it may distort our conclusions about female and white-nonwhite wage differentials, the age-earning the rate of return and the rate of depreciation of human cap the determinants of labor-force participation.

The economics of information has been established by now as part of economic analysis. The effect of the scarcity of informa the dispersion of prices and wages, the quality of goods, frictional un employment, and the microeconomics of inflation has been widely discussed in the literature.[1] However, surprisingly little has been written on the implications of search (in particular, job search) for the estimation of the wage function and its ramifications in such cases as the estimation of the determinants of labor-force participation, age-earning profiles,

This study is part of the NBER's project on economic and social measurement funded by the National Science Foundation. The paper has not undergone the full critical review accorded the National Bureau's studies. I am greatly indebted to H. Gregg Lewis and Robert T. Michael for their extensive notes. I benefited from the suggestions of Gary Becker, Stanley Diller, Victor Fuchs, James Heckman, Thomas Juster, and Esther Samuel. Finally, I would like to thank Margo Faier for her devoted research assistance.

[1] The economics of information dates back to Stigler's seminal papers (1961, 1962) in the early 1960s. The theory gained renewed popularity in the early 1970s in the context of the microeconomic analysis of the Phillips curve (Gronau 1971; McCall 1970; Mortensen 1970; Phelps 1970). For other applications and a survey of the theory see Nelson (1970) and Rothschild (1971).

[*Journal of Political Economy*, 1974, vol. 82, no. 6]

© 1974 by The University of Chicago. All rights reserved.

rates of return and rates of depreciation of human capital, degree of discrimination, etc.[2]

Given a wage-offer distribution, the parameters of the observed wage distribution depend on the intensity of search. The lower a person's wage demands, the greater the chance of his finding an acceptable job, but the lower wage he expects to receive and the wider the dispersion of acceptable wages around their mean. On the other hand, the job seeker may opt for a more ambitious search strategy, raising his minimum wage demand and consequently increasing the risk of remaining unemployed, but also increasing the expected wage and decreasing the dispersion of the acceptable offers.

Much of the discussion in labor economics concerning labor-force participation, wages, and earnings centers on the wage-offer distribution. However, the empirical validation of the theory is based on the observed wage distribution. The implicit identification of wage offers with observed wages is particularly suspect when those employed constitute only a fraction of the total population. In this case, the observed distribution represents only one part of the wage-offer distribution, while the other part is rejected by the job seekers as unacceptable. Thus, the traditional estimation procedures may involve certain biases when applied to the secondary labor groups— married women, teenagers, and the aged.

This paper attempts to point out some of these biases and suggests a method for their correction. In the next section we outline a simplified search model. The implications of this model for the investigation of labor-force participation, discrimination, and the rates of return and rates of depreciation of human capital are discussed in the third section. In the fourth section I describe an estimation procedure for the wage-offer distribution using a simplified set of assumptions. The paper closes with a discussion of some results obtained using the new method.

Correcting for what I call the "selectivity bias," I show that traditional measures underestimate the rate of return to human capital and its rate of depreciation when applied to married women. These measures tend to overstate the white-nonwhite wage differential, but to understate the wage-offer differentials between males and females and between women with and without young children. The use of data on the average wage of working women yields upward-biased estimates of the effect of wages on the labor-force participation of married women, as well as overestimates (in absolute terms) of the age and education effects.

The Search Model

Economic literature contains a large variety of search models. These models vary in their description of the search behavior and their assumptions about the job seeker's time horizon and his prior knowledge of market conditions. Some authors assume that the job seeker decides ahead of time on the number of searches he will undertake, while others adopt a sequential search model. Some assume that the wage-offer distribution is well known, while others postulate that the job seeker revises his assumptions about the distribution as a consequence of the search. Some allow for an infinite time horizon, while others restrict the time horizon.

---

[2] One of the few exceptions is Lewis (1971).

For simplicity, let us assume that the job seeker is unemployed, that he adopts a sequential search strategy, that he has perfect knowledge of the job-offer distribution (but that he does not know what is the wage associated with any specific vacancy), that he has an infinite time horizon, and that job-hopping is prohibitively expensive.[3] The job seeker is faced by a stream of job offers. Let it be assumed, for simplicity, that these offers arrive at a uniform rate, and let us define the time interval in such a way that the job seeker samples one offer per period. Firms differ in their search costs (e.g., the cost of ascertaining the worker's marginal productivity) and consequently may offer different wages W to the same job seeker, where W consists of nonpecuniary as well as pecuniary returns. *Let $f(W)$ denote the density function of W.*

The job hunter decides on a wage W_n^* to distinguish between those
wage offers that he deems acceptable and those that he rej
search process ends as soon as the job seeker receives an offer th
W_n^*. Let W^0 denote the job seeker's price of time at home; then
no wage offer that falls short of W^0, that is, $W_n^* \geq W^0$. In g
job seeker determines W_n^* so that any acceptable job offer will
of an income stream not inferior to the one he expects if he
his search. Let R_n denote the present value of a \$1.00 wage
accepted in period n,

$$R_n = \sum_{t=n+1}^{\infty} (1 + r)^{-t} = (1 + r)^{-n}/r, \tag{1}$$

where it is assumed that the rate of discount r is constant. Let I_n be the present value of the income stream the job seeker expects to receive if he continues his search beyond period n (both R_n and I_n are discounted to period 0); then $R_n W_n^* \geq I_n$, that is, $W_n^* \geq I_n/R_n$.

Assuming decisions are made at the end of the period [i.e., offers accepted in period n start yielding returns only in period $(n + 1)$], the job seeker enjoys in period n an income of $W^0 - C$, where C denotes the costs of search, which include both direct costs (e.g., advertisement, employment-agency fees, transportation) and indirect ones (e.g., the value of leisure foregone owing to the search).

---

[3] Thus, I rule out a search strategy that calls for the job seeker to accept the first job offered, continue his search while on the job, and switch to a better job as soon as it becomes available. This kind of strategy is inconsistent with voluntary unemployment. Admittedly, this model, first proposed in the literature by McCall (1970) and Mortensen (1970), is based on somewhat simplified assumptions. The use of a more sophisticated approach would not have affected our major conclusion that there is a positive correlation between the initial wage demands and the accepted wage.

Given the wage-offer distribution, $f(W)$, and the asking wage, W_n^*, there exists a probability of P_n that the job seeker will accept a job in period n, where

$$P_n = \text{Prob}\,(W_n \geq W_n^*) = \int_{W_n^*}^{\infty} f(W)\,dW. \tag{2}$$

An acceptable job will yield on the average a wage of E_n, where

$$E_n = E(W_n \mid W_n \geq W_n^*) = \frac{1}{P_n}\int_{W_n^*}^{\infty} Wf(W)\,dW. \tag{3}$$

If the offer received in period n is found to be unacceptable [an event whose assigned probability is $(1 - P_n)$], the job seeker can still look forward to an expected income stream of I_n. A job seeker embarking on search in period n can, therefore, expect an income stream of

$$I_{n-1} = (1 + r)^{-n}(W^0 - C) + P_n R_n E_n + (1 - P_n)I_n, \tag{4}$$

where r is the rate of discount. Given the infinite time horizon, the constant rate of discount, constant costs of search (C), constant current earnings (W^0), and a constant wage-offer distribution $[f(W)]$, the job seeker in period $(n + 1)$ is faced by the same conditions facing the job seeker in period n. Hence, the same optimum strategy that was employed in period n will be adopted in period $(n + 1)$, that is, $W_n^* = W_{n+1}^* = W^*$.[4] Consequently, the present value of the returns to search in period n (discounted to period n) should equal the present value of the returns to search in period $(n + 1)$ [discounted to period $(n + 1)$]. Put differently,

$$I_{n-1} = (R_{n-1}/R_n)I_n = (1 + r)I_n. \tag{5}$$

Moreover, the probability of acceptance (P_n) and the average acceptable wage (E_n) remain constant over time.

Inserting equations (5) and (1) in equation (4) and given that $W_n^* =$

---

[4] For a more rigorous proof of this statement, see McCall (1970) and Mortensen (1970). This conclusion does not hold if the job seeker's time horizon is finite, if imperfections in the capital market result in an increase in r as the search proceeds, if there are increasing costs of search, if current earnings (e.g., marginal utility of leisure) diminish, or if the job seeker changes his evaluation of $f(W)$ as the result of the search (see Gronau 1971).

W^*, $P_n = P$, and $E_n = E$ for every n, the asking wage equals[5]

$$W^* = \frac{1}{r + P}[r(W^0 - C) + PE]. \tag{6}$$

In other words, the minimum acceptable wage will be such as to equate the cost of search plus foregone earnings with the net returns from search:

$$C + (W^* - W^0) = P\frac{E - W^*}{r}. \tag{7}$$

The lower boundary of W^* is W^0. Let P^0 be the probability that the wage offer exceeds W^0, and E^0 the average acceptable wage when the job seeker's wage demands are confined to W^0. The job seeker quits his search when $W^0 > I_n/R_n$, that is, when

$$C > P^0\frac{E^0 - W^0}{r}. \tag{8}$$

For example, in the case of married women, when their price of time in home activities is sufficiently high relative to their market productivity (i.e., if $W^0 > E^0 - rC/P^0$), the woman will decide to stay out of the labor force altogether.

Other things being equal, an increase in the job seeker's pr
(W^0) reduces the foregone earnings associated with the reje
wage offer ($W^* - W^0$) and thus increases the job seeker's wag
though at a lower rate than the increase in the price of time.

$$dW^* = \frac{r}{r + P}dW^0 \leq dW^0.$$

The increase in wage demands reduces the probability of employment,

$$dP = -f(W^*)\,dW^* = -\frac{r}{r + P}f(W^*)\,dW^0 < 0, \tag{10}$$

but increases the average acceptable wage,

$$dE = \frac{f(W^*)}{P}(E - W^*)\,dW^* = \frac{r}{r + P}\frac{f(W^*)}{P}(E - W^*)\,dW^0 > 0. \tag{11}$$

---

[5] Inserting (5) and (1) in (4),

$$W_n^* = I_n/R_n = r(1 + r)^n[(1 + r)^{-(n+1)}(W^0 - C) + P_{n+1}R_{n+1}E_{n+1} + (1 - P_{n+1})I_{n+1}].$$

Since $W_n^* = W_{n+1}^* = W^*$, $P_n = P_{n+1} = P$, and $E_n = E_{n+1} = E$, then

$$\begin{aligned} W^* &= (1 + r)^{-1}[r(W^0 - C) + PE + (1 - P)(I_{n+1}/R_{n+1})] \\ &= (1 + r)^{-1}[r(W^0 - C) + PE + (1 - P)W^*]. \end{aligned}$$

Rearranging terms, one obtains eq. (6).

Moreover, since the returns to entry into the labor force [i.e., $P^0(E^0 - W^0)/r$] are a diminishing function of W^0, and since the cost of search (C) may increase with W^0, this change increases the tendency to abstain from entering the labor market.

Given W^0, an upward shift in the wage-offer distribution, that is, an increase in the mean wage-offer distribution, μ_W [other parameters of $f(W)$ remaining constant], increases the returns to search and hence the probability of labor-force participation and wage demands. However, the adjustment in wage demands lags behind the shift in $f(W)$,

$$dW^* = \frac{P}{r + P} d\mu_W, \tag{12}$$

resulting in an increase in the probability of employment,

$$dP = f(W^*)(d\mu_W - dW^*) = \frac{r}{r + P} f(W^*)\, d\mu_W > 0. \tag{13}$$

The increase in the wage demands intensifies the effect of the shift of $f(W)$ on the expected acceptable wage offer. The average acceptable wage increases, though at a slower rate than the shift in μ_W:

$$\begin{aligned} dE &= d\mu_W + \frac{f(W^*)}{P} (E - W^*)\,(dW^* - d\mu_W) \\ &= \left[1 - \frac{r}{r + P} \frac{f(W^*)}{P} (E - W^*)\right] d\mu_W. \end{aligned} \tag{14}$$

When the shift in the wage-offer distribution is accompanied by an increase in the price of time, the tendency to enter the labor force increases as long as $d\mu_W > dW^0$. Both the shift in $f(W)$ and the increase in W^0 tend to increase wage demands. Assuming $dW^0 < d\mu_W$, wage demands increase at a slower rate than the shift in $\mu_W (dW^* < d\mu_W)$, resulting in both an increase in the probability of employment and an increase in the average acceptable wage (the latter, however, changes more slowly than μ_W).

The Wage-Offer Distribution versus the Observed Wage Distribution: The Selectivity Bias

The model described in the previous section analyzes the job seeker's search strategy. One can design a somewhat similar model to describe the employer's search policy. There are still some unresolved problems: how the two strategies interact and what process determines simultaneously the rate of unemployment, the vacancy rate, and the distribution of wages. The formulation of such a model is clearly beyond the scope of this paper.

At the risk of being unrigorous I shall therefore adopt a partial-equilibrium approach. I assume that the wage-offer distribution is given and is not affected by the job seeker's strategy.[6] Under this strong assumption, the observed wage distribution is a subset of the wage-offer distribution, that is, that part of the distribution acceptable to the job seeker.

Wage data constitute a prime source of information for the estimation of labor supply and demand, the determinants and the effects of the investment in human capital (e.g., schooling, on-the-job training, health, migration), and the analysis of occupational choice. Objections have been raised to the indiscriminate use of such data, which usually reflect only the average pecuniary returns to a person's labor before the deduction of taxes. Hence, the data do not allow a distinction between the marginal wage and the average wage, do not account for nonpecuniary returns (i.e., psychic income) and the value of on-the-job training, and are overstated to the extent that work involves direct costs (e.g., commuting) and that the supply of labor is affected by after-tax wages. Several ingenious methods have been devised to overcome these shortcomings.

One bias that seems to have escaped economists' attention is that introduced by the search process. Given the wage-offer distribution, the bolder a person's search strategy (i.e., the higher one's asking wage W^*), the higher the wage he expects to accept. Thus, the observed wage tribution has to be adjusted for differences in the search poli

Nowhere is this bias more serious than in the case of th labor-force groups. These groups are characterized by partia tion in the labor force, indicating that portions of the wag tribution faced by these groups are considered too low to be The *observed wage distribution* is a *truncated section of the* distribution, and its parameters depend on the parameters of as well as on the truncation point W^*. Unless variations in W^* are corrected for, one is bound to obtain biased estimates of the parameters of $f(W)$.

Let us consider two groups of women who face the same wage-offer distribution, $f(W)$; have the same costs of search, C; and the same discount rate. Let it also be assumed that both groups have the same price-of-time distribution, except that the mean price of time (μ_{W^0}) of one group is higher than that of the second [let these distributions be denoted by $g(W^0 - K)$, where $K_1 > K_2 > 0$ and $d\mu_{W^0} = dK$]. Since both groups face the same opportunities, $f(W)$, and have the same costs of search, they have the same critical price of time, that is, the price that distinguishes between women who participate and those who do not participate in the labor force [$\overline{W}^0 = E^0 - (rC/P^0)$]. Their labor-force participation rate

---

[6] Thus I rule out the possibility that the employer is aware of the job seeker's search strategy and adapts his wage offer accordingly.

θ equals

$$\theta = \text{Prob}\,(W^0 > \overline{W}^0) = \int_{-\infty}^{\overline{W}^0} g(W^0 - K)\, dW^0. \tag{15}$$

This rate is inversely related to the mean price of time (and to K):

$$d\theta = -g(\overline{W}^0)\, d\mu_{W^0}. \tag{16}$$

Given the wage-offer distribution, the price of time W_i^0 determines the asking wage W_i^*, which in turn determines the average acceptable wage E_i. Given a sufficiently long period of search (a large enough n), the average wage of the labor-force participants in a certain group equals

$$\bar{E} = E(E \mid W^0 < \overline{W}^0) = \frac{1}{\theta}\int_{-\infty}^{\overline{W}^0} Eg(W^0 - K)\, dW^0. \tag{17}$$

It can be shown that when $g(W^0)$ is logconvex (a property that holds for most of the functions we have in mind), a leftward shift in the price-of-time distribution (i.e., a decline in K) results in a reduction of the wage demands and the average wage, $d\bar{E}/d\mu_{W^0} > 0$.[7]

For example, let us assume two groups of married women, say, whites and nonwhites, with the same market characteristics (e.g., education and work experience). If there were no discrimination, and if we ignore differences in psychic incomes, both groups should face the same wage-offer distribution, $f(W)$ (fig. 1). But, if an average white woman places on her time at home a higher value (W^0) than an average nonwhite because of her husband's higher earnings and the existence of other sources of income, her wage demands (W^*) will be higher, and her realized average wage (E) will be higher, though the probability that she participates in the labor force may be lower. A comparison of the average

---

[7] Let us define the density function

$$h_K(W^0) = \begin{cases} g(W^0 - K)/\theta_K & \text{where } W^0 \le \overline{W}^0, \\ 0 & \text{where } W^0 > \overline{W}^0, \end{cases}$$

where $\theta_K = \text{Prob}(W^0 \le \overline{W}^0)$. The likelihood ratio of $h_K(W^0)$ depends solely on the likelihood ratio of $g(W^0 - K)$:

$$h_{K_1}(W^0)/h_{K_2}(W^0) = (\theta_{K_2}/\theta_{K_1})[g(W^0 - K_1)/g(W^0 - K_2)],$$

since $(\theta_{K_2}/\theta_{K_1})$ is a constant independent of W^0 (it depends on $\overline{W}^0$, K_1, and K_2). Thus, $h_K(W^0)$ has a monotone likelihood ratio if $g(W^0 - K)$ has one, a necessary and sufficient condition for which is that $-\log g$ is convex (Lehmann 1959, p. 330). Since E has been shown to be a nondecreasing function of W^0, if $h_K(W^0)$ has a monotone likelihood ratio in W^0, then

$$\bar{E} = \int_{-\infty}^{\infty} Eh_K(W^0)\, dW^0$$

is a nondecreasing function of K (p. 74).

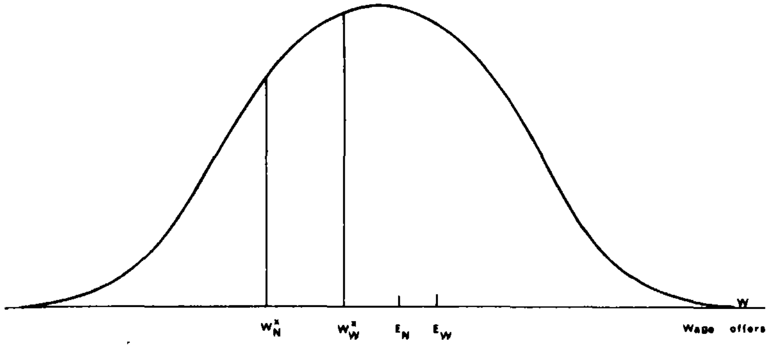


FIG. 1

wages of working women belonging to the two groups (E_W and E_N) may lead to the conclusion that nonwhite women are discriminated against, while in effect discrimination is nonexistent.

On the other hand, if the two groups have the same price-of-time distribution, the same C and r, and the same wage-offer distribution except for a shift factor, then by (11) and (14) the difference in the critical price of time ($\overline{W}^0$) between the two groups is equal to the differen between their mean wage offers. We shall expect the group higher mean wage offer to have the higher participation rate [$d\theta/d\overline{W}^0 = g(\overline{W}^0) > 0$]. The increase in the average wage, ho short of the increase in the mean wage offer,

$$dE = [1 - (dE/d\mu_{W^0})]\, d\mu_W < d\mu_W.$$

Thus, if it is assumed that education affects market productivity (hence the wage-offer distribution) but does not affect nonmarket productivity and the price of time, women with higher education are expected to have a higher participation rate. The average wage of these women, however, increases more slowly than their mean wage offers. These conclusions will hold, even if we remove the assumption that education does not affect nonmarket productivity, as long as it is assumed that its effect on market productivity exceeds that of nonmarket productivity (see fig. 2).

Wage comparisons of whites and nonwhites and of women with different levels of education are two of the cases that may be affected by what may be called the "selectivity bias." In the remaining part of this section I shall discuss a few other cases where this kind of bias may be prevalent.

Male-female wage differentials.—More than 95 percent of males in the prime age groups (25–55) participate in the labor force,[8] while the rate of participation of females, and in particular married women, almost

---

[8] U.S. Bureau of the Census (1971), p. 21.

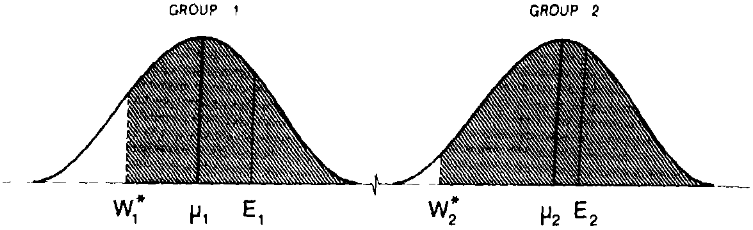


FIG. 2

never exceeds 55 percent.[9] The difference in participation can be explained by the lower wage-offer distribution facing women, and probably by their higher value of time in the absence of market opportunities. Both factors tend to increase the difference between the average acceptable wage and the mean wage offer.

It has been estimated (Fuchs 1971) that female average hourly earnings adjusted for color, schooling, age, city size, marital status, class of worker, and length of work trip constituted in 1959 only two-thirds of the hourly earnings of nonfarm males. Given the "selectivity bias," it seems that this figure underestimates the "true" male-female wage differentials.

Age-wage profiles.—Labor-force participation rates of males vary very little in the 25–54 age group, decline somewhat (to a level of about 80 percent) in the 55–64 group, and fall very steeply (to 25 percent) thereafter. Female participation rates have a bimodal distribution with peaks in the 20–24 and 45–54 age groups.[10] These variations in the participation rates may bias the estimates of the age-wage profiles.

Age-wage profiles (or age-earning profiles) are used frequently to estimate the rate of depreciation of human capital (Rosen 1972) and the rates of return to education and on-the-job training. Retirement may introduce a bias in the estimate of the rate of depreciation and obsolescence of human capital. The direction of the bias is, however, indeterminate. It depends on who is the first to retire, those with the high wage offers or those with the low wage offers. The increase in nonhuman capital with age may increase the shadow price of time, which will rise more for those who benefited over time from the higher wage offers, so that they will be the first to retire. On the other hand, the margin between the market wage and the home wage is smallest for those at the lower tail of the wage distribution. Any deterioration in wage offers may therefore make them leave the labor force.

Likewise, as argued above, intereducational variations in the participation rates may bias the estimated rate of return on investment in human capital of women. It is well established that women's labor-force

[9] Ibid., p. 24.
[10] Ibid.

participation increases with education. Put differently, an increase in education shifts the wage-offer distribution by more than it increases the woman's price of time at home (the latter increases either because of the effect education may have on her home productivity and/or because of the increased probability that she is married to a man with higher earnings). Ignoring the effect education may have on the dispersion of the wage-offer distribution, the difference between the observed average rate (E) and the mean wage-offer distribution shrinks with education.[11] The difference between the observed wages of working women in two education groups therefore tends to understate the difference between their mean wage offers and the estimated rate of return to education.

It has been argued (Michael and Lazear 1971) that the woman's wage rate may be affected by the number and age composition of her children. The existence of young children increases the demand for the wife's time at home and her price of time and reduces her tendency to participate in the labor force. Leaving the labor market may accelerate the depreciation of the woman's market-oriented skills and shift her wage-offer distribution downward. However, the decline in the wage-offer distribution does not necessarily result in a decline in the observed wage of working mothers. If the increase in these women's wage demands is sufficiently large, it will offset the decline in the wage offers and resu
an increase in the observed average wage.[12]

The determinants of labor-force participation.—Labor-force
increases with the mean of the wage-offer distribution and
W^0. A prerequisite for the estimation of the determinants of
is knowledge of the distribution of the price of time at hom
1973; Gronau 1973*b*; Lewis 1971). However, this informa
cient unless supplemented by knowledge of the mean
Replacing the latter by the average observed wage yields only in

---

[11] An increase in the standard deviation of the wage-offer distribution leads to an increase in wage demands:

$$dW^* = \frac{P}{r+P}\bar{X}\,d\sigma_W > 0, \qquad \text{where } \bar{X} = (E - \mu_W)/\sigma_W.$$

Its effect on the probability of acceptance and the average acceptable wage seems, however, to be indeterminate:

$$dP = \frac{r}{r+P} f(W^*)[W^0 - (\mu_W + C)]\frac{d\sigma_W}{\sigma_W},$$

$$dE = \left[\bar{X} - \frac{f(W^*)}{P}\left(X^* - \frac{P}{r+P}\bar{X}\right)(E - W^*)\right] d\sigma_W,$$

where $X^* = (W^* - \mu_W)/\sigma_W$.

[12] By eqq. (11) and (14) the average acceptable wage increases if

$$dW^0 > \left[1 - \frac{r+P}{r}\frac{P}{f(W^*)}\frac{1}{E - W^*}\right] d\mu_W.$$

estimates of the parameters of the labor-force participation function, unless one knows the relationship between the average wage and the mean wage offer ($d\bar{E}/d\mu_W$).

It was shown that an increase in the mean wage offer (μ_W) leads to an increase in the observed wage (E), but that this increase is smaller than the original shift in μ_W. Differences in the observed wages of different, say, education groups therefore tend to understate the difference between the mean wage offers and to overstate the sensitivity of participation rates of changes in wages. Furthermore, since the correlation between μ_W and E is not perfect, the misspecification of the wage variable introduces errors of measurement into the equation that may also bias the other estimates of the determinants of labor-force participation.

The Estimation of the Wage-Offer Distribution

The observed wage distribution is only one section of the wage-offer distribution, namely, those wages that exceed the asking wage W^* where the price of time of the job seeker falls short of $\bar{W}^0$. To prevent the kind of selectivity bias discussed in the preceding section, the original wage-offer distribution must be generated. The derivation of the wage-offer distribution from the truncated observed distribution is a difficult task even if the point of truncation (i.e., W^*) is known, and it verges on the impossible when W^* and its determinants, that is, the price of time, the costs of search, and the rate of discount, are not known. In the following section I adopt an oversimplified version of the search model. This section is therefore merely illustrative, serving, it is hoped, as a base point for future research.

Let it be assumed that the job seeker is aware of the best offer he can attain, W,[13] and let his policy be to stay out of the labor force if his price of time W^0 exceeds his best offer W and otherwise to accept the wage offer W. Given these oversimplified assumptions, the participation rate is

$$\theta = \text{Prob}\,(W > W^0) = \text{Prob}\,(w = W - W^0 > 0) = \int_0^\infty g(w)\,dw, \quad (18)$$

where $g(w)$ is the density function of the differential $w = W - W^0$. The average wage of working women is

$$E = E(W \mid W > W^0) = \frac{1}{\theta}\int_{-\infty}^{\infty}\int_{W^0}^{\infty} Wh(W, W^0)\,dW\,dW^0, \quad (19)$$

where $h(W, W^0)$ denotes the joint density distribution of W and W^0.

---

[13] Note that from this point on W denotes the best wage offer and $f(W)$ denotes the best-wage-offer distribution rather than the wage-offer distribution.

Let it be assumed that $h(W, W^0)$ is a bivariate normal density function,

$$h(W, W^0) = (2\pi\sigma_W\sigma_{W^0}\sqrt{1-\rho^2})^{-1}$$
$$\exp\left[-\frac{1}{2(1-\rho^2)}(X^2 - 2\rho XY + Y^2)\right], \tag{20}$$

where $X = (W - \mu_W)/\sigma_W$, $Y = (W^0 - \mu_{W^0})/\sigma_{W^0}$, and ρ is the correlation coefficient between W and W^0. The differential w has a normal distribution $w \sim N(\mu, \sigma^2)$ where

$$\mu = \mu_W - \mu_{W^0} \qquad \text{and} \qquad \sigma^2 = \sigma_W^2 + \sigma_{W^0}^2 - 2\rho\sigma_W\sigma_{W^0}.$$

The participation rate is

$$\begin{aligned} \theta &= (2\pi\sigma^2)^{-1/2}\int_0^\infty \exp\left[-\frac{1}{2}\left(\frac{w-\mu}{\sigma}\right)^2\right]dw \\ &= (2\pi)^{-1/2}\int_{-\mu/\sigma}^\infty \exp\left(-\frac{1}{2}Z^2\right)dZ, \end{aligned} \tag{21}$$

where Z is a standardized normal variable $Z = (w - \mu)/\sigma$. The average wage is[14]

$$E = \mu_W + (1 - \beta)(\sigma_W^2/\sigma)\bar{X}, \tag{22}$$

where β is the regression coefficient of W^0 on $W(=\rho\sigma_{W^0}/\sigma_W$

$$\bar{X} = g(Z = -\mu/\sigma)/\theta.$$

Let it be assumed that the (best) wage offers depend on race, age, education, and number of children, but not on h income.[15] The sample can be classified by race, age, educat of children, and husband's income. Given the rate of participati each cell, the value of the ratio $-(\mu/\sigma)$, that is, the value satisfying Prob $(Z > -\mu/\sigma) = \theta$, can be generated from the normal tables.[16] Given the value of $-(\mu/\sigma)$, one can use the same tables to derive the value of $g(Z = -\mu/\sigma)$, and hence the value of $\bar{X}$. Assuming that σ_{W^0} does not change with income, one can regress within a race, age, education, number-of-children group j for different income groups

$$E = a_{0j} + a_{1j}\bar{X}, \tag{24}$$

where a_{0j} = est (μ_{Wj}) and a_{1j} = est $(1 - \beta)(\sigma_{Wj}^2/\sigma_j)$. Moreover, if one adopts the stronger assumption that σ_W^2 does not vary with race, age, education, and number of children (i.e., that σ is constant), one can estimate for the whole sample

---

[14] See Appendix.

[15] I ignore the possible correlation between the husband's income and the wife's wage offers that may arise because of the positive correlation in their natural ability.

[16] For example, if $\theta = 0.16 \Rightarrow -\mu/\sigma = 1$, $\theta = 0.50 \Rightarrow -\mu/\sigma = 0$, $\theta = 0.84 \Rightarrow -\mu/\sigma = -1$.

$$E = \sum a_{0j}D_j + a_1\bar{X}, \tag{25}$$

where D_j is a dummy variable representing race, age, education, number-of-children group j, and where $a_{0j} = \text{est}\,(\mu_{Wj})$.

The Data and the Results

To estimate the wage-offer distribution, I used the 1960 census one one-thousandth sample. I focused on married women, with the sample consisting of 26,530 women belonging to urban primary families, spouse present. Assuming that the mean of the wage-offer distribution is a function of race, age, education, and number of children, the data were subclassified by these characteristics: 2 race groups (white, Negro) × 4 age groups (below 30, 30–39, 40–49, 50+) × 4 education groups (elementary school, high school, college, and graduate studies) × 3 groups for number of children below the age of 6 (0, 1, 2+). Since it was assumed that the wife's price of time in the absence of market opportunities is, in addition to these factors, a function of income, the data were further divided by income excluding wife's earnings (12 groups: less than $2,000, 2,000–2,999, 3,000–3,999, . . . , 9,000–9,999, 10,000–14,999, 15,000–19,999, and 20,000+).[17] This subclassification yielded 1,152 cells that are the basic observations of my sample.

For each cell I computed two statistics: (*a*) the labor-force participation rate of women belonging to the cell (i.e., the percentage of women working or looking for work during the week preceding the census), and (*b*) the average hourly wage rate of working women belonging to the cell. The 1960 census did not contain any direct evidence on the hourly wage rate, and I had to do with an imperfect substitute, defining the hourly wage as the ratio of the woman's 1959 earnings divided by the product of the number of weeks she worked in 1959 and the number of hours she worked during the week preceding the census. Finally, to rule out the possibility of negative mean wage offers, the assumption of bivariate normality was replaced by the assumption of bivariate log-normality. Thus, instead of computing the usual arithmetic mean wage rate, I computed the arithmetic mean of the natural logarithm of the wage rates. The dependent variable E in equation (25) therefore denotes the natural logarithm of the geometric mean of the hourly wage rates.

Equation (25) was estimated for four groups: whites with no children under 6 years old, whites with one child under 6, whites with two or more children under 6, and nonwhites with no children under 6 (the other

---

[17] This classification is very similar, but not identical, to the one I used before (Gronau 1973*b*).

TABLE 1

ESTIMATION OF MEAN WAGE OFFER, $E = a_{0j}D_j + a_1\bar{X}$

EDUCATION AND AGE	WHITES (CHILDREN UNDER SIX)						NONWHITES (ZERO CHILDREN UNDER SIX)	
	0		1		2+			
	a	*t*	*a*	*t*	*a*	*t*	*a*	*t*
0–8 years of schooling:								
<30	−.2163	3.29	−.2265	1.35	−.4656	0.24	0.2939	1.28
30–39	.0019	2.51	−.0585	0.55	−.2939	0.48	−0.3852	3.16
40–49	−.0364	4.13	−.1458	0.65	*	...	−0.4815	3.91
50+	−.1935	5.82	*	...	*	...	−0.3720	2.50
9–12 years of schooling:								
<30	.1344	0.31	−.0589	1.01	−.4223	0.05	0.0619	0.20
30–39	.1466	2.74	.0307	1.34	−.4158	2.45	0.0325	0.24
40–49	.1211	0.77	.2494	1.04	*	...	−0.0785	0.83
50+	.0212	2.71	*	...	*	...	−0.3953	2.27
13–16 years of schooling:								
<30	.4930	5.81	.1436	0.86	−.0510	2.01	0.5857	1.90
30–39	.3566	3.69	.2283	1.31	.0733	2.14	0.6204	2.61
40–49	.4695	6.57	.1426	0.36	*	...	0.6077	2.30
50+	.3998	4.22	*	...	*	...	0.5211	1.65
17+ years of schooling:								
<30	.7761	3.63	.6084	1.54	*	...	1.3195	[illegible]
30–39	.9313	5.39	.5642	1.37	*	...	[illegible]	
40–49	.8078	6.38	*	..	*	...		
50+	.9054	5.43	*	...	*	...		
$\bar{X}$	.3577	6.01	.3150	1.91	.4539	2.05		
Adjusted R^2	...	0.60	...	0.05	...	0.14		

* No observations in the cell.

two groups are too small to allow estimation). The results are [illegible] in table 1.[18] All the regressions are significant, and all the coefficients of $\bar{X}$ are positive as expected (they are significant in three out of the four cases, with the exception of the regression for nonwhites).

Given the estimates of a_{0j}, one can compute the median of the log-normal best-wage-offer distribution [$\mu_{Wj} = \exp(a_{0j})$]. These estimates are presented in table 2 and figure 3 for the group of white women with

---

[18] Instead of eq. (25), I estimated

$$\bar{X} = a'_{00} + \sum_{j=1}^{15} a'_{0j}D_j + a_1\bar{X}, \tag{25'}$$

where the basic group (i.e., the subscript 00) is the group of women 30–39 years old with high school education. To derive the estimators of the mean wage offers (a_{0j} in eq. [25]), one has to compute $a_{0j} = a'_{00} + a'_{0j}$. These estimates are presented in table 1. However, I have not recalculated the t-coefficients. Thus, the t_j coefficients ($j = 1, \ldots, 15$) presented in table 1 reflect the significance of the differential between the mean wage offers of the jth group and the base group (i.e., they serve to test the hypotheses H_0: $\mu_{Wj} = \mu_{W0}$), while the t-coefficient for the base group measures how significantly different is the mean wage offer of this group from zero (i.e., H_0: $\mu_{W0} = 0$).

TABLE 2

GEOMETRIC AVERAGE WAGE AND THE MEDIAN WAGE OFFER BY AGE AND EDUCATION: WHITE MARRIED WOMEN WITH NO YOUNG CHILDREN

Age	Elementary School	High School	College	Postgraduate
Geometric average wage:				
< 30	1.12	1.45	1.93	2.35
30–39	1.37	1.53	1.90	3.02
40–49	1.37	1.54	2.11	2.59
50+	1.39	1.56	2.17	3.20
Median wage offer:				
< 30	0.81	1.14	1.64	2.17
30–39	1.02	1.16	1.43	2.54
40–49	0.96	1.13	1.60	2.24
50+	0.82	1.02	1.49	2.47

no young children, together with the original data of the geometric means of the hourly wage rates.

Figure 3 demonstrates two of the salient features of the age-wage profiles of married women. The first of these, often discussed in the literature, is their flatness. Thus, except for the wages of women graduates, there is very little change in the average wage of working married women over their life cycle.[19] The second feature, one rarely commented on, is the fact that the profiles show no tendency to slope downward as a result of negative net investment in human capital.[20]

The replacement of the average wage of working women by the imputed median wage offers has very little effect on the slope of the early parts of the age profiles, but reverses the slope of the later segment (the exception is again the graduate group). It seems that a married woman faces her peak wage offer at the age of 30–40 when she has elementary or high school education, and at the age of 40–50 when she is a graduate.[21]

Sixteen observations are too small a sample to derive any strong conclusions. Still, at the risk of seeming foolhardy, I could not resist the temptation to apply to the observations in table 2 some of the techniques devised and applied with such success by Jacob Mincer (1974) to earnings

---

[19] This flatness may be partly due to a cohort effect and does not necessarily imply that women entering the labor force now can expect no increase in their wage rates.

[20] Of course, this may be merely due to the crude aggregation of the wage data. The expected downward-sloping portion of the age-wage profile might have emerged had I used a more refined age classification. However, it is noteworthy that Arleen Leibowitz, who fitted a quadratic function to data of weekly wages of married women who are full-time workers, found that the wage rates tend to decrease only after 43 years of experience when the woman is a high school graduate, and after 60 years of experience when she has 13 years of schooling or more (Leibowitz 1972, p. 43). This would imply that the peak of the age-wage profile is reached at the age of 61 if the woman is a high school graduate and at the age of 83 if she finished college (the regression coefficients in the case of elementary school are insignificant).

[21] The negatively sloping portion of the wage-offer profiles may, of course, be due purely to intercohort differences.

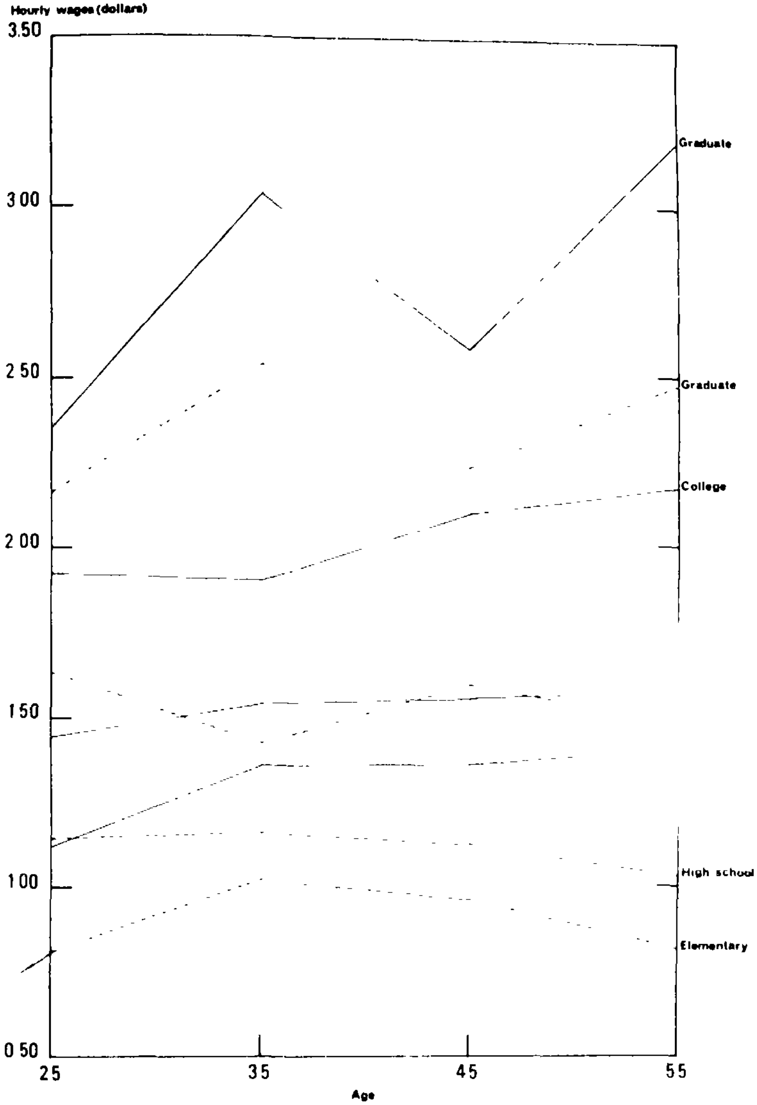


FIG. 3.—Solid line = average wage; dashed line = wage offer.

of prime-age males. Assuming that schooling and on-the-job training involve no direct costs, and that the ratio of "time-equivalent" investment in on-the-job training declines linearly with age (Mincer 1974, chap. 3), I estimated the function

$$E = b_0 + b_1 S + b_2 T + b_3 T^2, \tag{26}$$

TABLE 3

ESTIMATION OF RATE OF RETURN TO EDUCATION AND EFFECT OF WORK EXPERIENCE

DEPENDENT VARIABLE	INTERCEPT	EXPLANATORY VARIABLE			ADJUSTED $\bar{R}^2$
		S	T*	T^2	
Average wage rate (E):					
Regression (1):					
b	−0.3901	0.0655	0.3048	0.2256	.80
t	2.97	7.72	0.42	0.13	...
Regression (2):					
b	−0.2273	0.0584	...	...	.76
t	2.26	7.01	...	...	...
Median wage offer (μ_W):					
Regression (1):					
b	−0.7429	0.0748	0.4356	−0.1587	.83
t	4.79	7.46	0.51	0.80	...
Regression (2):					
b	−0.7986	0.0797	...	...	.84
t	7.35	8.86	...	...	...

* T is measured in 100-year units.

where S denotes the number of years of schooling (it was assumed that these are 8, 12, 15, and 18 for the four education groups, respectively), and T denotes the number of years of working experience (defined as $T = \text{age} - S - 8$).[22] To correct for heteroscedasticity, each of the observations was weighted by the number of working women belonging to that age-education group.[23] The results of equation (26) are compared with the estimates of a similar regression equation in which the estimated median wage offer is the dependent variable (see table 3).

The strong multicollinearity between the variable T and its square (the simple correlation coefficient is .97) undermines the reliability of the estimated effect of market experience.[24] Keeping the shakiness of the evidence in mind, it is still interesting to compare the coefficients of T and T^2. Thus, while the coefficients of both T and T^2 in the regression of the average wage rate are positive, implying an upward-sloping age-wage profile, the coefficient of T^2 is negative when the dependent variable is the imputed wage offers, as one would expect. The wage offers increase at an ever-diminishing rate, reflecting the decline in investment in human capital over the life cycle. An absolute decline in wage offers sets in after 25–30 years of experience ($-b_2/b_3 = 27$).

---

[22] Age is assumed to be 26, 35, 45, and 55 in the four age groups, respectively.

[23] The major conclusions of this analysis would not have changed had I used a different weighting scheme, e.g., estimating an unweighted regression or weighting each observation by the total number of women belonging to that group.

[24] One should not conclude from the low t-values associated with T and T^2 that length of work experience does not play any role in the determination of wage rates of married women. The omission of T^2 from eq. (26) results in the coefficient of T becoming significantly different from zero (see also Leibowitz 1972, p. 43).

TABLE 4

Geometric Average Wage and Median Wage Offer by Race and Number of Young Children

	Whites (Children under Six)			Nonwhites (Zero Children under Six)
	0	1	2+	
1. Average wage of working woman:				
Dollars per hour	1.59	1.53	1.46	1.00
Relative*	1.00	0.96	0.92	0.63
2. Average potential wage rate:				
Dollars per hour	1.56	1.53	1.36	0.95
Relative	1.00	0.98	0.87	0.61
3. Standardized potential wage rate:				
Dollars per hour	1.56	1.81	1.67	1.02
Relative	1.00	1.16	1.07	0.65
4. Standardized median wage offer:				
Dollars per hour	1.09	1.08	0.72	0.93
Relative	1.00	0.99	0.66	0.85

* Relative wage is the wage of group i divided by the corresponding wage of white women with no young children.

Assuming that education has the same effect on the productivity of time in all its uses (e.g., work in the market, work at home, education and leisure), the coefficient of years of schooling in equation (26) measures the rate of return to formal education (Mincer 1974, chap. 3). The comparison of the coefficient of S in the regression of the wage rate with the mean wage offer confirms our prior expectations. Th observed wage rather than the imputed wage offer tends t ward the estimated rate of return to schooling. The extent increases when the experience variables (T and T^2) are the regression.

The data on the average wage rates in the other three groups (women with young children and nonwhites with no young children), and consequently the other three regressions, show no clear systematic pattern of behavior, and I have not attempted to apply equation (26) to these groups,[25] but have made do with a comparison of the average wage and wage offers of the women belonging to these groups. This comparison is contained in table 4.

The first line of table 4 describes the (geometric) average wage of working women belonging to the four groups investigated in our study. There are only slight differences in average wages among the three groups of white women, regardless of the number of young children, and there exists a substantial gap between the wage enjoyed by white women and that by nonwhites: the latter is less than two-thirds of the former. As the second line of this table indicates, this pattern is little affected by using as weights the total number of women belonging to a cell rather than the number of working women belonging to the cell.

---

[25] Any attempt of this kind in the case of white women with young children would have proved futile in the face of the large number of empty cells in these groups.

The average wage of each group naturally depends on the group's age and educational composition. To standardize for differences in composition, one has to use a uniform weighting scheme. The weights used to compute the "standardized potential wage rate" (line 3) are the number of white women without young children belonging to the age and education group. The isolation of the age and education effect only slightly reduces the white-nonwhite differential (from 0.61 to 0.65) but changes the wage hierarchy of the white women. Thus, while the unstandardized figures give an edge to women without young children (the wage of women with two young children is 87 percent of the wage of women without young children), this relationship is reversed when one accounts for age and education differences between the groups (the ratio becomes 1.07). These figures thus appear to contradict the hypothesis that children result in accelerated depreciation of the mother's market skills.[26]

The replacement of the average wage by the standardized imputed median wage offer (line 4) leads to significant changes in the observed wage pattern. The effect of children on the mother's wage offers turns out to be a prominent feature of the wage structure. The wage a white mother with two young children can expect to get in the market is only two-thirds of that of a woman belonging to the same age-education group who does not have young children. (On the other hand, one young child does not seem to affect his mother's wage offers.) Furthermore, the removal of the selectivity bias seems to remove more than one-half of the white-nonwhite differential (nonwhites' wage offers are 85 percent of those of whites), calling for major revisions in the evaluation of the importance of discrimination as a factor affecting the wage structure of nonwhite women.

On the other hand, it seems that the estimate of the male-female differential must be revised upward. The median wage offer of married women is only about 70 percent of the standardized average wage. Since the male labor-force participation rate is close to 100 percent, the data on their wages are hardly affected by the selectivity bias ($\bar{X}$ is very close to zero). The comparison of wage offers of males and females would therefore show a substantially larger differential than the 3 to 2 ratio quoted earlier in this paper.

Finally, it has been argued (Lewis 1971) that using the average wage of working women rather than their mean wage offer results in an upward

---

[26] This statement has to be qualified since the results presented in table 4 are not standardized for the number of children older than 6.

TABLE 5

DETERMINANTS OF HOUSEWIVES' VALUE OF TIME: AN INTERACTION MODEL

EXPLANATORY VARIABLES	POTENTIAL WAGE = AVERAGE WAGE: Probit Coefficients	*t*-Scores	Marginal Effect on μ_{WO}	POTENTIAL WAGE = MEDIAN WAGE OFFER: Probit Coefficients	*t*-Scores	Marginal Effect on μ_{WO}
Constant	−1.894	1.58	...	0.870	1.44	...
Annual income*	−0.850	3.91	9.5	−0.823	3.86	17.1
Age:†						
< 30	0.159	0.46	−1.8	−0.346	1.18	7.1
≥ 50	−0.870	3.24	9.7	−0.249	0.85	5.1
Education:†						
Elementary	0.081	0.19	−0.9	−0.072	0.18	1.5
College	−2.138	2.43	23.8	−1.135	1.82	23.3
Husband's age‡	−0.296	2.60	3.3	−0.206	1.84	4.2
Husband's education:†						
Elementary	0.418	2.10	−4.6	0.410	2.08	−8.4
College	−0.126	0.58	1.4	−0.128	0.59	2.6
Number of children < 3:						
Elementary	−0.861	1.31	9.6	−0.868	1.32	17.8
High school	−1.077	4.20	12.0	−1.071	4.20	22.0
College	−1.880	3.49	20.9	−2.012	3.71	41.4
Number of children 3–5:						
Elementary	0.114	0.20	−1.3	0.106	0.18	−2.2
High school	−0.768	3.53	8.5	−0.783	3.61	16.1
College	−0.443	0.99	4.9	−0.536	1.19	11.0
Number of children 6–11:						
Elementary	−0.186	0.79	2.1	−0.215	[illegible]	
High school	−0.382	2.71	4.3	−0.389	[illegible]	
College	−0.364	1.22	4.0	−0.390	[illegible]	
Number of children 12–17:						
Elementary	0.275	1.05	−3.1	0.235	[illegible]	
High school	−0.291	1.78	3.2	−0.256	[illegible]	
College	0.337	0.96	−3.8	0.369	1.06	
Potential hourly wage ($)	8.982	3.35	...	4.866	3.29	...
Likelihood-ratio test	169.0	...	...	165.0	...	...
Degrees of freedom	21	...	...	21	...	...

SOURCE.—Gronau (1973*a*).
* $10,000.
† Dummy variable.
‡ 10 years.

bias in the estimated effect of wages on labor-force participation. I have elsewhere (Gronau 1973*a*) used the results presented in table 2 to estimate the price-of-time distribution and the implicit participation function of American married women. Employing probit analysis, I examined the effect of the woman's and her husband's age and education, family income, and the number and age composition of the children on her price of time, that is, her labor-force participation (I allowed the effect of children to differ according to their mother's education). The results of this analysis are reproduced in table 5, where one regression contains the (geometric) average wage rate of working women as an explanatory variable, and the other regression uses our estimate of μ_W.

These results clearly confirm our expectations. The introduction of the wage offers as an explanatory variable considerably reduces the estimated wage effect. The misspecification of the wage variable also affects the estimated effects of other variables, notably the wife's age and education.[27] We have seen earlier that the observed wage of working women belonging to the 50+ age group considerably overstates the wage offers faced by their nonworking counterparts. The decline in labor-force participation witnessed in that group is therefore attributed to age, while essentially it is merely a wage effect. Likewise, changes in the observed wage tend to understate the changes in the wage offers associated with an increase in education. Using the average wage of working women tends to overplay the direct effect of education on labor-force participation.[28]

The data on the hourly wage rate used in our estimates are far from perfect. One might prefer a long-term measure of participation, say, annual rather than weekly participation. Data problems are enhanced by a set of admittedly very strong assumptions, notably the assumption of homoscedasticity of the price of time and the wage-offer distribution. Finally, the estimation model is based on a search model that is clearly suboptimal. Thus, one should regard the results of this section as no more than an empirical exercise. One should strive to overcome these shortcomings by using a better-suited body of data, by assuming a different joint distribution of W and W^0 that calls for weaker assumptions, or by focusing on a different statistic (e.g., the standard deviation of the wage

---

[27] I owe this point to Robert E. Hall. Note that the effect of the misspecification of the wage variable on the estimated effect of any variable on the housewife's price of time is indeterminate. This effect is measured as the ratio of the probit coefficients of that variable and the coefficient of the wage variable. The effect of the misspecification in this case therefore depends on the direction and magnitude of the bias in both probit coefficients. For example, though the education coefficient is biased (in absolute terms) upward, the effect of education on the woman's price of time remains unchanged when E is replaced by μ_W.

[28] Let the "true" function be

$$\theta = \delta_0 + \delta_1 \mu_W + \delta_2 S, \qquad \delta_1 > 0, \quad \delta_2 < 0,$$

where S denotes education. The replacement of μ_W by E,

$$\theta = C_0 + C_1 E + C_2 S,$$

results in biased estimators of δ_i,

$$\begin{aligned} \text{Plim } (C_1) &= \delta_1 \cdot \delta_{\mu E \cdot S}, \\ \text{Plim } (C_2) &= \delta_2 + \delta_1 \cdot \delta_{\mu S \cdot E}. \end{aligned}$$

If, as was claimed, $\delta_{\mu E \cdot S} > 1$ and $\delta_{\mu S \cdot E} < 0$,

$$\text{Plim } (C_1) > \delta_1 \quad \text{and} \quad |\text{Plim } (C_2)| > |\delta_2|.$$

of working women).[29] Nevertheless, it seems, at least to me, that the evidence collected here is a strong enough warning that comparisons of wage data should take selectivity bias into account.

Appendix[30]

Let

$$
\begin{aligned}
X &= (W - \mu_W)/\sigma_W,\\
Y &= (W^0 - \mu_{W^0})/\sigma_{W^0},\\
w &= W - W^0,\\
\mu &= \mu_W - \mu_{W^0},\\
\sigma^2 &= \sigma_W^2 + \sigma_{W^0}^2 - 2\rho\sigma_W\sigma_{W^0},\\
Z &= (w - \mu)/\sigma,\\
\text{and } Z_0 &= -(\mu/\sigma).
\end{aligned}
$$

Define

$$\varepsilon = (W^0 - \mu_{W^0}) - \beta(W - \mu_W) = \sigma_{W^0}(Y - \rho X),$$

where $\beta = \rho(\sigma_{W^0}/\sigma_W)$ is the regression coefficient of W^0 on W. The mean value of ε is zero and its variance is $\sigma_\varepsilon^2 = \sigma_{W^0}^2(1 - \rho^2)$. Define the standardized variable $\lambda = \varepsilon/\sigma_\varepsilon$, and let

$$\lambda^* = (\varepsilon - \mu)/(1 - \beta)\sigma_W = (\lambda\sigma_\varepsilon - \mu)/\sqrt{(\sigma^2 - \sigma_\varepsilon^2)} = A$$

where $A = -\mu/\sqrt{(\sigma^2 - \sigma_\varepsilon^2)}$ and $B = \sigma_\varepsilon/\sqrt{(\sigma^2 - \sigma_\varepsilon^2)}$.

When $\beta < 1$,

$$
\begin{aligned}
\theta &= \text{Prob}\ (W > W^0)\\
&= \text{Prob}\left\{\frac{1}{(1 - \beta)\sigma_W}[\beta(W - \mu_W) + (1 - \beta)(W - \mu_W) - (W^0 - \mu_{W^0}) + \mu] > 0\right\}\\
&= \text{Prob}\left\{\frac{(1 - \beta)(W - \mu_W)}{(1 - \beta)\sigma_W} > \frac{[(W^0 - \mu_{W^0}) - \beta(W - \mu_W)] - \mu}{(1 - \beta)\sigma_W}\right\}\\
&= \text{Prob}\ [X > (\varepsilon - \mu)/(1 - \beta)\sigma_W = \lambda^*].
\end{aligned}
$$

---

[29] A set of data that seems very alluring in this context is the *National Longitudinal Survey: Survey of Work Experience of Females 30–44* (the "Ohio Survey"), which contains information on the asking wage of nonworking women. Analysis of the variance of the observed wage distribution (S_W^2) may yield further insights into the problems discussed above, since it can be shown that if W and W^0 are normally distributed,

$$S_W^2 = \sigma_W^2[(1 - \beta)^2(\sigma_W^2/\sigma^2)(\bar{X} + Z)\bar{X}].$$

Assuming σ_W and σ_{W0} are constant, one can regress S_W^2 on $\bar{X}(\bar{X} + \bar{Z})$ to obtain independent estimates of these parameters.

[30] This Appendix is based on the extensive comments of H. Gregg Lewis.

The conditional expectation

$$\begin{aligned} E &= E(W \mid W > W^0) \\ &= E(\mu_W + X\sigma_W \mid W > W^0) = \mu_W + \sigma_W E(X \mid X > \lambda^*) = \mu_W + \bar{X}'\sigma_W. \end{aligned}$$

Since X and Y are normally distributed, ε and λ, which are linear combinations of X and Y, are normally distributed. Furthermore, by definition X and ε—and in turn, X and λ—are orthogonal. Hence,

$$h'(X, \lambda) = (2\pi)^{-1} \exp\left[-\tfrac{1}{2}(X^2 + \lambda^2)\right].$$

Thus,

$$\begin{aligned} \bar{X}' &= E(X \mid X > \lambda^*) = (2\pi\theta)^{-1} \int_{-\infty}^{\infty} \int_{\lambda^*}^{\infty} X \exp\left[-\tfrac{1}{2}(X^2 + \lambda^2)\right] dX\, d\lambda \\ &= (2\pi\theta)^{-1} \int_{-\infty}^{\infty} \exp\left[-\tfrac{1}{2}(\lambda^2 + \lambda^{*2})\right] d\lambda \\ &= (2\pi\theta)^{-1} \int_{-\infty}^{\infty} \exp\left\{-\tfrac{1}{2}[A^2 + 2AB\lambda + (1 + B^2)\lambda^2]\right\} d\lambda \\ &= (2\pi\theta)^{-1} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left[(1 + B^2)\left(\lambda + \frac{AB}{1 + B^2}\right)^2 \right.\right. \\ &\qquad\qquad\qquad\qquad \left.\left. + \left(A^2 - \frac{(AB)^2}{1 + B^2}\right)\right]\right\} d\lambda \\ &= (2\pi\theta)^{-1} \exp\left(-\frac{1}{2}\frac{A^2}{1 + B^2}\right) \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}(1 + B^2)\left(\lambda + \frac{AB}{1 + B^2}\right)^2\right] d\lambda \\ &= \frac{1}{\theta\sqrt{\{2\pi(1 + B^2)\}}} \exp\left(-\frac{1}{2}\frac{A^2}{1 + B^2}\right). \end{aligned}$$

Inserting

$$1 + B^2 = \sigma^2/(\sigma^2 - \sigma_\varepsilon^2) = \sigma^2/(1 - \beta)^2\sigma_W^2$$

and

$$A^2/(1 + B^2) = (-\mu/\sigma)^2 = Z_0^2,$$

it is found that

$$\bar{X}' = [(1 - \beta)\sigma_W/\sigma][\exp(-\tfrac{1}{2}Z_0^2)/\theta\sqrt{(2\pi)}] = [(1 - \beta)\sigma_W/\sigma]\bar{X},$$

where

$$\bar{X} = \exp(-\tfrac{1}{2}Z_0^2)/\theta\sqrt{(2\pi)} = g(Z_0)/\theta.$$

When $\beta > 1$,

$$\begin{aligned} \theta &= \mathrm{Prob}\,(W > W^0) = \mathrm{Prob}\,(X < \lambda^*), \\ \bar{X}'' &= E(X \mid X < \lambda^*) = -[(\beta - 1)\sigma_W/\sigma]\bar{X} = \bar{X}'. \end{aligned}$$

Hence, in general,

$$E = \mu_W + (1 - \beta)(\sigma_W^2/\sigma)\bar{X}.$$

References

Ben-Porath, Yoram. "Labor-Force Participation Rates and the Supply of Labor." *J.P.E.* 81, no. 3 (May/June): 697–704.

Fuchs, Victor R. "Differences in Hourly Earnings between Men and Women." *Monthly Labor Rev.* 94 (May 1971): 9–15.

Gronau, Reuben. "Information and Frictional Unemployment." *A.E.R.* 61, no. 3, pt. 1 (June 1971): 290–301.

———. "The Effect of Children on the Housewife's Value of Time." *J.P.E.* 81, no. 2, pt. 2 (March/April 1973): S168–99. (*a*)

———. "The Intra-Family Allocation of Time: The Value of the Housewives' Time." *A.E.R.* 63 (September 1973): 634–51. (*b*)

Lehmann, E. L. *Testing Statistical Hypotheses.* New York: Wiley, 1959.

Leibowitz, Arleen S. "Women's Allocation of Time to Market and Nonmarket Activities: Differences by Education." Ph.D. dissertation, Columbia Univ., 1972.

Lewis, H. Gregg. "On Income and Substitution Effect in Labor Force Participation." Mimeographed. Chicago: Univ. Chicago, 1971.

McCall, J. J. "Economics of Information and Job Search." *Q.J.E.* 84, no. 1 (February 1970): 113–26.

Michael, Robert, with Lazear, Edward P. "On the Shadow Price of Children." Paper presented at the meetings of the Econometric Society, New Orleans, December 1971.

Mincer, Jacob. *Schooling, Experience, and Earnings.* New York: Nat. Bur. Econ. Res., 1974.

Mortensen, Dale T. "Job Search, the Duration of Unemployment, and
Phillips Curve." *A.E.R.* 60 (December 1970): 847–62.

Nelson, Phillip. "Information and Consumer Behavior." *J.P*
(March/April 1970): 311–29.

Phelps, E., ed. *Microeconomic Foundations of Employment Theory.* New
1970.

Rosen, S. "Measuring the Obsolescence of Knowledge." Mimeog

Rothschild, Michael. "Models of Market Organization with Imper
tion: A Survey." Discussion Paper no. 224, Harvard Inst. Econ.
bridge, Mass.: Harvard Univ., 1971.

Stigler, George J. "The Economics of Information." *J.P.E.* 69, no. 3 (June 1961): 213–25.

———. "Information in the Labor Market." *J.P.E.* 70, no. 5, pt. 2 (October 1962): 94–105.

U.S., Bureau of the Census. *Statistical Abstract of the United States 1971.* Washington: Government Printing Office, 1971.

Comments on Selectivity Biases in Wage Comparisons

H. Gregg Lewis

University of Chicago

I. Search-related Selectivity Bias in Wage Comparisons between Homogeneous Groups

These comments on selectivity biases in wage comparisons stem from Gronau's paper on the subject in this issue of the *Journal* and assume that the reader is familiar with his paper.

Consider a group of persons who differ in their home price of
but who are alike in having the same discount rate r, sear
distribution $f(W)$ of market wage offers W. I shall call
"homogeneous" group. Let $\bar{E}$ be the mean market wage of
group who, after a "sufficiently long period of search," are
the market; μ_W is the mean and σ_W^2 the variance of their
offer distribution; μ_{W^0} is the mean and $\sigma_{W^0}^2$ the variance of th
of time W^0 in the group; and θ is the group's labor-force part
rate. Then the search theory used by Gronau implies that:

$$\begin{aligned} &0 \leq a \equiv \bar{E} - \mu_W = a(\mu, r, C, \sigma_W, \sigma_{W^0}) \leq E^0 - \mu_W; \\ &0 \leq \theta = \theta(\mu, rC, \sigma_W, \sigma_{W^0}) \leq 1; \qquad \mu \equiv \mu_W - \mu_{W^0}; \end{aligned} \tag{1}$$

where E^0 is, as defined in his paper, the mean acceptable wage offer of a person at the margin of indifference between searching and not searching—that is, a person whose home price of time W^0 is equal to the critical price of time $\bar{W}^0$.

The bias a measures the extent to which the mean market wage $\bar{E}$ of employed persons in the group overstates the mean μ_W of their common wage-offer distribution. The bias a is zero, of course, if the dispersion σ_W of the wage-offer distribution is zero; indeed, a may be zero even if σ_W

[*Journal of Political Economy*, 1974, vol. 82, no. 6]
© 1974 by The University of Chicago. All rights reserved.

exceeds zero.[1] In general, however, the bias a exceeds zero—$\bar{E}$ exceeds μ_W—and is a function of the excess μ of the mean μ_W of the wage-offer distribution over the group mean μ_{W^0} of the home price of time, the discount rate r, search cost C, the dispersion σ_W of the wage-offer distribution, and the dispersion σ_{W^0} of the home price of time in the group. Similarly, the labor-force participation rate θ for the group depends upon μ, rC, σ_W, and σ_{W^0}.

Write the equation in (1) for the bias a for each of two homogeneous groups (denoted by subscripts 1 and 2) that differ only with respect to μ, r, C, σ_W, and σ_{W^0} and subtract the equation for group 2 from that for group 1:

$$\begin{aligned} A_{12} &\equiv a_1 - a_2 = (\bar{E}_1 - \bar{E}_2) - (\mu_{W_1} - \mu_{W_2}) \\ &= a(\mu_1, r_1, C_1, \sigma_{W_1}, \sigma_{W_1^0}) - a(\mu_2, r_2, C_2, \sigma_{W_2}, \sigma_{W_2^0}). \end{aligned} \tag{2}$$

In general, A_{12} is not zero. That is, search tends to produce a selectivity bias in wage comparisons so that the observed difference $\bar{E}_1 - \bar{E}_2$ between the groups in the mean wage of employed persons does not measure accurately the true difference $\mu_{W_1} - \mu_{W_2}$ in the mean of the wage-offer distribution.

What are the implications of Gronau's search model for the sign of the selectivity bias A_{12}? Suppose that the two groups differ only in having different means μ_W of the wage-offer distributions they face or in having different means μ_{W^0} of the home price of time. In his search theory, the two means μ_W and μ_{W^0} enter the bias a-function and the labor-force participation θ-function for a group as a single variable $\mu = \mu_W - \mu_{W^0}$.[2] He shows that if the density function $g(W^0)$ of the home price of time W^0 for the group is logconvex, then the bias a is a negatively inclined function of μ, which implies that if the two groups differ only with respect to μ, the selectivity bias A_{12} has the sign of

$$\mu_2 - \mu_1 = (\mu_{W_2} - \mu_{W_1}) - (\mu_{W_2^0} - \mu_{W_1^0}).$$

However, in general $\mu_2 - \mu_1$ cannot be observed directly. Indeed, I think that it is unlikely that $\mu_2 - \mu_1$ could be observed directly without also observing $\mu_{W_2} - \mu_{W_1}$; but then there would be no need to be concerned about the selectivity bias A_{12}. Thus, the sign of $\mu_2 - \mu_1$ must be obtained indirectly. Gronau shows that if $0 < \theta < 1$, the labor-force participation rate θ for the group is a positively inclined function of μ, so that the sign of $\mu_2 - \mu_1$ is that of $\theta_2 - \theta_1$. Therefore, if only μ varies between the groups, the sign of the selectivity bias A_{12} is that of $\theta_2 - \theta_1$: the extent to which the observed mean wage $\bar{E}$ overstates the mean μ_W of the wage-offer distribution is greater for the group with the smaller participation rate θ.

---

[1] Let the minimum of the wage-offer distribution be $\mu_W - \delta$, where $\delta \geq 0$. Then if $\delta \leq rC$, the critical price of time $\bar{W}^0$ in Gronau's paper is equal to $\mu_W - rC$ and is less than or equal to the minimum wage offer $\mu_W - \delta$. Then a is zero even though δ and σ_W may exceed zero.

[2] This result of Gronau's search theory can be said to be a consequence of his assumption that search cost C is the same given for each person in the group and would not hold if it had been assumed instead that search cost C for each person is, say, a linear function of his home price of time W^0.

The preceding argument may be put in the following way: solve the labor-force participation rate equation in (1) for μ and substitute the result for μ in the bias equation, obtaining a new bias function in which θ replaces μ:

$$\begin{aligned} 0 \leq a \equiv \bar{E} - \mu_W &= a[\mu(\theta, rC, \sigma_W, \sigma_{W0}), r, C, \sigma_W, \sigma_{W0}] \\ &= \alpha(\theta, r, C, \sigma_W, \sigma_{W0}) \leq E^0 - \mu_W; \end{aligned} \tag{3}$$

where the partial derivative of α with respect to θ is negative if $0 < a < E^0 - \mu_W$, $0 < \theta < 1$, and $g(W^0)$ is logconvex.[3] That this negative partial derivative of the bias a with respect to the labor-force participation rate θ does not imply that the selectivity bias A_{12} has the sign of $\theta_2 - \theta_1$, unless r, C, σ_W, and σ_{W0} are the same for both homogeneous groups, is perhaps so obvious that it need not be said. Yet I fear that some readers of Gronau's paper may misread him as saying that in general the selectivity bias A_{12} has the sign of $\theta_2 - \theta_1$. I say this for two reasons. First, in the theoretical section of his paper he discusses in detail only the relation of selectivity bias and labor-force participation to the mean μ_W of the wage-offer distribution and to the home price of time (W^0 and μ_{W0}) and for the most part ignores r, C, σ_W, and σ_{W0}.[4]

---

[3] Some readers may find it useful to verify that if the density function of the home price of time W^0 in a homogeneous group is rectangular over the interval $\mu_{W0} - D^0/2 \leq W^0 \leq \mu_{W0} + D^0/2$, where $D^0 \geq 0$ and the wage-offer distribution is binomial—either a wage $\mu_W + qD$ with probability $p = 1 - q$, or a lower wage $\mu_W - pD$ with probability q, where $D \geq 0$, then the bias function has the following form:

$$\begin{aligned} &0 \leq a = \frac{qD(r + p)(pD - rC)}{rpD^0\theta} \leq qD; 0 \leq \theta = \frac{1}{2} + \frac{\mu + qD - rC/p}{D^0} \leq 1; \\ &D = \sigma_W/\sqrt{pq}; D^0 = \sigma_{W0}\sqrt{12}; \end{aligned} \tag{23}$$

where, if any of the indicated limits is violated, a or θ, as the case may be, should be set equal to the limit. The form of the bias function and that of the participation-rate function in this example are misleading, of course, in their simplicity. However, the signs of the partial derivatives are not misleading.

[4] If $0 < a < E^0 - \mu_W$ and $0 < \theta < 1$, then Gronau's search theory implies that in general: (*a*) The bias function α is a negatively inclined function of θ, if $g(W^0)$ is logconvex. But it is also negatively inclined function of r, C, and σ_{W0}, and it is a positively inclined function of σ_W. (*b*) The labor-force participation rate θ is a positively inclined function not only of $\mu \equiv \mu_W - \mu_{W0}$, but also of σ_W; a negatively inclined function of rC; and a function of σ_{W0} that has the shape of an inverted *U*.

Second, in some of his illustrations of wage comparisons in which search-related selectivity bias may arise, he appears to use the proposition that A_{12} has the sign of $\theta_2 - \theta_1$ as though it held in general.

Furthermore, Gronau's search model assumes that search cost C is a given constant as long as a person is not employed in the market, but becomes prohibitively great as soon as he accepts a market wage offer. Assume instead that search cost C is the same given for a person whether he is or is not employed in the market. Then a person will gain from search as long as his current wage is less than the critical price of time $\bar{W}^0$. Therefore, in this case the mean acceptable wage E of all searchers in a group is E^0; hence, "given a sufficiently long period of search," the mean wage $\bar{E}$ of employed persons in a homogeneous group is simply E^0 and the bias a is $E^0 - \mu_W$, which is independent of θ.[5]

My most serious reservation regarding the search model used by Gronau is that it takes the wage-offer distribution as a datum with a dispersion that may be large enough (see n. 1) to imply an observed wage distribution that differs substantially from the wage-offer distribution. But surely the wage-offer distribution is not a datum. In particular, it seems plausible to suppose that when market forces have brought the labor market into equilibrium, differences between the two distributions will not exist. Thus, I am not convinced that in equilibrium the search bias a is not zero.

II. Participation-related Selectivity Bias in Wage Comparisons between Heterogeneous Groups

Because our knowledge of search, labor-force participation, and wage-offer distributions is incomplete, to put it mildly, in the actual wage comparisons that we make, the groups whose wages are compared are heterogeneous rather than homogeneous. That is, within each of the groups there is dispersion not only in the home price of time, but also in wage-offer distributions and perhaps in search costs and discount rates, as well. In the theoretical section of his paper, Gronau does not deal explicitly with selectivity biases in wage comparisons between heterogeneous groups, though much of what I have to say on this subject in this and the next section of these comments is derived from the model that he uses for his empirical work.

Suppose, as before, that within a group there is dispersion in the home price of time W^0, but none in either search cost C or discount rate r. On the other hand, I now assume that there are differences within the group in the wage-offer distributions faced by the members of the group. However, to simplify, I assume that the wage-offer distributions differ among persons only with respect to the mean μ_W. There will then be some joint distribution of W^0 and μ_W within the group. For purposes of illustration I assume that W^0 and μ_W are normally correlated—that is, the joint density function of W^0 and μ_W is the bivariate normal density function.

---

[5] In the example given in n. 3, the bias a is either zero or is equal to qD.

Since, by assumption, all members of the group have the same value of rC and wage-offer distributions that differ only with respect to the mean μ_W, it follows from Gronau's search theory that the excess $\bar{W}^0 - \mu_W$ of the critical price of time $\bar{W}^0$ over the mean μ_W of the wage-offer distribution is the same for all persons in the group. That is,

$$\bar{W}^0 = \mu_W + \eta, \tag{4}$$

where η, the same for all persons in the group, depends upon rC and moments of the wage-offer distribution of higher order than the mean. Therefore, $\bar{W}^0$ also is normally correlated with W^0.

As before, let μ_{W^0} be the group mean and $\sigma^2_{W^0}$ the group variance of the home price of time W^0. Similarly, $\mu_{\bar{W}^0}$ is the group mean and $\sigma^2_{\bar{W}^0}$ the group variance of the critical price of time $\bar{W}^0$; and $\sigma_{\bar{W}^0W^0}$ is the group covariance of $\bar{W}^0$ and W^0. Now define the participation criterion w:

$$w \equiv \bar{W}^0 - W^0; \tag{5}$$

it pays a person to search (participate in the labor market) if and only if w exceeds zero. The group mean μ_w of w is $\mu_{\bar{W}^0} - \mu_{W^0}$ and the corresponding variance is $\sigma^2_w = \sigma^2_{\bar{W}^0} + \sigma^2_{W^0} - 2\sigma_{\bar{W}^0W^0}$. In what follows it will be a bit simpler to use as the participation criterion the standardized value Z of w:

$$Z \equiv \frac{w}{\sigma_w} + \gamma = \frac{w - \mu_w}{\sigma_w}; \qquad \gamma \equiv -\frac{\mu_w}{\sigma_w}; \tag{6}$$

where a person gains from search (participates) if and only if $Z > \gamma$. The group mean μ_Z of Z of course is zero, and the variance σ^2_Z of Z is unity.

Since Z is a linear function of the normally correlated variables $\bar{W}^0$ and W^0, it is normally distributed. Therefore, the labor-force participation rate θ for the group is

$$\theta = \frac{1}{\sqrt{2\pi}} \int_{\gamma}^{\infty} e^{-(1/2)Z^2}\, dZ = \theta(\gamma). \tag{7}$$

That is, $\theta = \theta(\gamma)$ depends only on $\gamma \equiv -\mu_w/\sigma_w$ and, indeed, from (7), θ is a declining function of γ. Notice that when θ is given, the value of γ can be read from a table of the areas under the normal curve; thus, γ is known when θ is known.

I will use the symbol M (appropriately subscripted) to stand for the mean of a variable among the labor-force participants in the group; the corresponding mean for the nonparticipants I will denote by N. Similarly, I will use the symbols m and n to stand for the variance of a variable among the participants and nonparticipants, respectively, in the group. Since Z is normally distributed with zero mean and unit variance and exceeds γ only for participants, it follows that

$$M_Z = \frac{e^{-(1/2)\gamma^2}}{\theta\sqrt{2\pi}}; \qquad N_Z = \frac{-e^{-(1/2)\gamma^2}}{(1-\theta)\sqrt{2\pi}}; \tag{8}$$

$$m_Z = 1 - M_Z(M_Z - \gamma); \qquad n_Z = 1 - N_Z(N_Z - \gamma).$$

Since γ is known when θ is known, $e^{-(1/2)\gamma^2}/\sqrt{2\pi}$ can be read from a table of the ordinates of the normal curve and then M_Z, N_Z, m_Z, and n_Z are easily calculated. $M_Z = -(1 - \theta)N_Z/\theta$ goes from infinity at $\theta = 0$ to zero at $\theta = 1$; m_Z goes from zero at $\theta = 0$ to unity at $\theta = 1$; and n_Z goes from unity at $\theta = 0$ to zero at $\theta = 1$.

The participation criterion Z is a linear function of the normally correlated variables $\bar{W}^0$ and W^0; therefore, each of the latter is normally correlated with Z. I denote the group covariances of $\bar{W}^0$ and W^0 with Z by $\sigma_{Z\bar{W}^0}$ and σ_{ZW^0} respectively. Since $w \equiv \bar{W}^0 - W^0 = \sigma_w(Z - \gamma)$, it follows that

$$\sigma_{Z\bar{W}^0} = \frac{\sigma^2_{\bar{W}^0} - \sigma_{\bar{W}^0W^0}}{\sigma_w}; \qquad \sigma_{ZW^0} = \frac{\sigma_{\bar{W}^0W^0} - \sigma^2_{W^0}}{\sigma_w}; \tag{9}$$

$$\sigma_w = \sigma_{Z\bar{W}^0} - \sigma_{ZW^0}.$$

The signs of the two covariances $\sigma_{Z\bar{W}^0}$ and σ_{ZW^0} in general are ambiguous, except that since $\sigma_w = \sigma_{Z\bar{W}^0} - \sigma_{ZW^0}$ cannot be negative, $\sigma_{Z\bar{W}^0}$ cannot be smaller algebraically than σ_{ZW^0}.

Moreover, since Z has a zero mean and unit variance and is normally correlated with each of $\bar{W}^0$ and W^0 with covariances $\sigma_{Z\bar{W}^0}$ and σ_{ZW^0}, respectively, then each of $\bar{W}^0$ and W^0 has a linear regression on Z within the group as follows:

$$\bar{W}^0 = \mu_{\bar{W}^0} + \sigma_{Z\bar{W}^0}Z + \lambda; \qquad W^0 = \mu_{W^0} + \sigma_{ZW^0}Z + \lambda; \tag{10}$$

where the residual λ is distributed normally and independently of Z with a zero mean and a variance σ^2_λ equal to

$$\sigma^2_\lambda = \sigma^2_i - \sigma^2_{Zi}, \qquad i = \bar{W}^0, W^0. \tag{11}$$

Therefore, it follows directly from (10) and (11) that

$$\left\{\begin{aligned} M_i - \mu_i &= \frac{1-\theta}{\theta}(\mu_i - N_i) = (1-\theta)(M_i - N_i) = \sigma_{Zi}M_Z; \\ m_i &= \sigma_i^2 - \sigma_{Zi}^2(1 - m_Z);\ n_i = \sigma_i^2 - \sigma_{Zi}^2(1 - n_Z); \end{aligned}\right\} \quad i = \bar{W}^0, W^0. \tag{12}$$

Assume that the labor-force participation rate θ for the group is neither zero nor unity. Then M_Z, $1 - m_Z$, and $1 - n_Z$ must exceed zero; moreover, since $\sigma_{Z\bar{W}^0} - \sigma_{ZW^0} = \sigma_w$ exceeds zero, both $\sigma_{Z\bar{W}^0}$ and σ_{ZW^0} cannot be zero. Then the mean $M_{\bar{W}^0}$ of the critical price of time $\bar{W}^0$ among labor-force participants in the group in general is a biased estimate, with a bias $\sigma_{Z\bar{W}^0}M_Z$ of ambiguous sign, of the true group mean $\mu_{\bar{W}^0}$ of $\bar{W}^0$. (The corresponding mean $N_{\bar{W}^0}$ for the nonparticipants also is a biased estimate, of course, of $\mu_{\bar{W}^0}$ with an oppositely signed bias). In the same way both M_{W^0} and N_{W^0} are biased estimates of μ_{W^0}. The two variances $m_{\bar{W}^0}$ and $n_{\bar{W}^0}$ of $\bar{W}^0$ among participants and nonparticipants, respectively, in general are both underestimates of the true group variance $\sigma^2_{\bar{W}^0}$ of $\bar{W}^0$ and similarly m_{W^0} and n_{W^0} underestimate $\sigma^2_{W^0}$. All of these biases are "selectivity" or "self-selection" biases. They are all related to group participation rates that differ both from zero and unity. They arise because the persons in the group choose whether to participate or not to participate on the basis of a wage criterion $w \equiv \bar{W}^0 - W^0$. In general they will be present in wage comparisons between heterogeneous groups irrespective of any search-related selectivity bias that also may be present.

Up to this point it has been unnecessary to say much about the underlying search theory. I now adopt the model used by Gronau, but to simplify I assume that search cost C is the same whether a person is or is not employed in the market. Then the mean acceptable wage E of each person in the group is E^0 and E^0 exceeds the wage-offer distribution mean μ_W of the person by the search-related bias a that depends only on rC and σ_W, which are the same for all persons in the group. That is,

$$E = E^0 = \mu_W + a, \tag{13}$$

where a is the same for all persons in the group. Return now to equation (4), solve for μ_W, and substitute the result into (13):

$$E = E^0 = a - \eta + \bar{W}^0. \tag{14}$$

Hence, "given a sufficiently long period of search," the mean market wage M_W of employed persons in the group is

$$M_W = a - \eta + M_{\bar{W}^0}. \tag{15}$$

Denote by μ_{μ_W} the group mean of the wage-offer distribution means μ_W. It follows from (4) that $\mu_{W^0} = \mu_{\mu_W} + \eta$. Substitute this result into (12) in the appropriate equation to obtain

$$M_{W^0} = \mu_{\mu_W} + \eta + \sigma_{ZW^0} M_Z. \tag{16}$$

Now substitute the right-hand side of (16) for M_{W^0} in (15)[6]

$$M_W - \mu_{\mu_W} = a + \sigma_{ZW^0} M_Z. \tag{17}$$

It follows from (17) that the mean wage M_W of employed persons in the group now contains two biases as an estimate of the group mean μ_{μ_W} of wage-offer distribution means μ_W: the search-related bias $a \geq 0$, as in the preceding section, and the new participation-related bias $\sigma_{ZW^0} M_Z$ of ambiguous sign.[7]

Now compare equation (17) with Gronau's equation (22), which is the key equation underlying his empirical work. In the comparison, make the following translation of symbols in which his symbol is given first and mine follows in parentheses:

$$E,\ (M_W); \qquad \mu_W,\ (\mu_{\mu_W}); \qquad \bar{X},\ (M_Z); \qquad \frac{\sigma_W^2(1-\beta)}{\sigma},\ (\sigma_{ZW^0}).$$

The two equations are identical in all respects except one: the search-related bias a in my equation (17) does not appear in his equation (22). Thus, in the model underlying his empirical work, he ignores the search-related bias emphasized in his theoretical section.

III. Other Selectivity Biases in Wage Comparisons between Heterogeneous Groups

In order to simplify exposition, I now abstract from search-related selectivity bias by assuming that the wage-offer distribution faced by each person has no dispersion. In this case the search economics becomes very simple. A person will search if and only if his home price of time W^0 is

---

[6] Had I adopted Gronau's search theory without modification, then eqq. (13)–(17) would have to be changed as follows. The mean acceptable wage E of a person then in general would be less than E^0, and $E - \mu_W$ would have been the same only for persons with the same value of the participation criterion Z. That is, (13) would have to be written

$$E = \mu_W + a', \tag{13'}$$

where a' now depends upon Z as well as rC and σ_W. The eqq. (14) and (15) would be changed only by replacing a by a'. Hence, (17) would be changed by replacing a by the search-related bias $M_{a'}$, the mean of a' among labor-force participants, where $M_{a'}$ is a negatively inclined function of θ.

[7] In Gronau's only illustration of a wage comparison involving a selectivity bias of indeterminate sign, namely, that of "age-wage profiles," the illustration deals with comparisons between heterogeneous groups. The cause of the indeterminacy is the participation-related bias $\sigma_{ZW^0} M_Z$ in (17).

less than the critical price of time $\bar{W}^0$. Moreover, if he searches at all, he will search only once, accepting the first offer of the single wage W in his wage-offer distribution. Therefore, the critical price of time $\bar{W}^0$ is simply the net wage $W - rC$ offered by the market—that is, the excess of the wage W offered by the market over search cost rC.

Return to equations (12), one of which is

$$M_{\bar{W}^0} - \mu_{\bar{W}^0} = (1 - \theta)(M_{\bar{W}^0} - N_{\bar{W}^0}) = \sigma_{Z\bar{W}^0}M_Z.$$

Thus, it might be said that the participation-related selectivity bias $\sigma_{Z\bar{W}^0}M_Z$ in $M_{\bar{W}^0}$ as an estimate of $\mu_{\bar{W}^0}$ arises because implicitly we are comparing the (net) wages of two groups, here participants and nonparticipants, who have selected the group to which they belong in part on the basis of the (net) wage offered to them by the market, so that neither group can be assumed to be a random sample of the two groups taken together.

When this language is used, it is immediately apparent that selectivity or self-selection biases may arise in numerous contexts other than search and labor-force participation. For example, it is common to estimate the return to schooling by comparing the wages of persons with different amounts of schooling (who may be alike with respect to such control variables as sex, color, and labor-market experience). Since individuals choose the amount of their schooling at least in part on the basis of the expected incremental return to incremental schooling, a selectivity bias may arise in the wage comparison even in the absence of any bias related to labor-force participation. Similarly, the return to geographic migration may be estimated by comparing the destination wages of migrants with the corresponding origin wages of nonmigrants. Here, too, there is likely to be a selectivity bias stemming from self-selection in migration.

The possibility of such self-selection bias in these and similar contexts has been widely recognized, and often diligent efforts have been made to control for it. Thus, the notion of selectivity bias is not new. However, I do want to suggest that Gronau's empirical model (his eq. [22] together with the eq. in his n. 26, or my eq. [12]) may be useful for estimating any such selectivity bias remaining after direct efforts to eliminate it have been made.

The example that I use here is that of migration. To simplify the analysis, I assume that there are only two places between which migration occurs. Consider then a group of persons who are alike with respect to place of origin, some of whom (nonmigrants) are in the origin place and others of whom (migrants) have migrated to the destination place. For each person there is a destination net wage $\bar{W}^0 = W - rC$, a corresponding origin wage W^0, and a net return $w \equiv \bar{W}^0 - W^0$ to migration. For each migrant the destination net wage $\bar{W}^0$, I assume, is observed, but the corresponding origin wage W^0 is not observed. Similarly, for each nonmigrant the origin wage W^0, I assume, is observed, but $\bar{W}^0$ is his unobserved, potential destination net wage. A person is a migrant if and only if his return to migration $w \equiv \bar{W}^0 - W^0$ exceeds zero, the fraction θ of the group who are migrants consists of all of those for whom $w > 0$, and hence Z exceeds $\gamma \equiv -\mu_w/\sigma_w = (\mu_{W^0} - \mu_{\bar{W}^0})/\sigma_w$.

I now use the symbols M and N, appropriately subscripted, to denote mean values for the migrant (M) and nonmigrant (N) components of the group; similarly, I use the symbols m and n to denote the corresponding variances. The problem is to estimate the (presumably positive) mean net return to migration $M_w = M_{\bar{W}^0} - M_{W^0}$ for migrants and the corresponding (presumably negative) mean net return to migration $N_w = N_{\bar{W}^0} - N_{W^0}$ for nonmigrants, using the observed data on the migration rate θ, the destination net wages $\bar{W}^0$ for migrants, and the origin wages W^0 for nonmigrants.

First notice that

$$\begin{aligned} M_{\bar{W}^0} - N_{W^0} &= (N_{\bar{W}^0} - N_{W^0}) + (M_{\bar{W}^0} - N_{\bar{W}^0}) \\ &= N_w + (M_{\bar{W}^0} - N_{\bar{W}^0}); \\ &= (M_{\bar{W}^0} - M_{W^0}) + (M_{W^0} - N_{W^0}) \\ &= M_w + (M_{W^0} - N_{W^0}). \end{aligned} \tag{18}$$

That is, the excess $M_{\bar{W}^0} - N_{W^0}$ of the mean destination net wage $M_{\bar{W}^0}$ of migrants over the mean origin wage N_{W^0} of nonmigrants is generally a biased estimate of both the return to migration M_w of migrants and that, N_w, of nonmigrants. The biases involved are given by equations (12):

$$M_{\bar{W}^0} - N_{\bar{W}^0} = \sigma_{Z\bar{W}^0}M_Z/(1 - \theta) \text{ and } M_{W^0} - N_{W^0} = \sigma_{ZW^0}M_Z/(1 - \theta).$$

Therefore, (18) can be written

$$M_{W^0} - N_{\bar{W}^0} = M_w + \sigma_{ZW^0}M_Z/(1 - \theta) = N_w + \sigma_{Z\bar{W}^0}M_Z/(1 - \theta). \tag{19}$$

Since σ_w is positive and equal to $\sigma_{Z\bar{W}^0} - \sigma_{ZW^0}$—see (9)—it follows that the bias in $M_{\bar{W}^0} - N_{W^0}$ as an estimate of N_w is algebraically greater than the corresponding bias in $M_{\bar{W}^0} - N_{W^0}$ as an estimate of M_w. However, both biases may be positive or both negative. Now take a weighted mean of these two equations, weighting the first by θ and the second by $(1 - \theta)$;

$$\begin{aligned} M_{\bar{W}^0} - N_{W^0} = {} & \theta M_w + (1 - \theta)N_w \\ & + M_Z[\theta\sigma_{ZW^0} + (1 - \theta)\sigma_{Z\bar{W}^0}]/(1 - \theta). \end{aligned} \tag{20}$$

$\theta M_w + (1 - \theta)N_w$ is simply the overall mean return to migration μ_w for the group and μ_w in turn is $-\sigma_w\gamma = (\sigma_{ZW^0} - \sigma_{Z\bar{W}^0})\gamma$. Hence, make this substitution in (20) and collect terms to obtain:

$$M_{\bar{W}^0} - N_{\bar{W}^0} = \sigma_{ZW^0}(\gamma - N_Z) + \sigma_{Z\bar{W}^0}(M_Z - \gamma), \tag{21}$$

where $N_Z = -\theta M_Z/(1 - \theta)$.

There are also the two equations from (12) for the variance $m_{\bar{W}^0}$ of $\bar{W}^0$ among migrants and the variance n_{W^0} of W^0 among nonmigrants:

$$m_{\bar{W}^0} = \sigma^2_{\bar{W}^0} - \sigma^2_{Z\bar{W}^0}(1 - m_Z); \quad n_{W^0} = \sigma^2_{W^0} - \sigma^2_{ZW^0}(1 - n_Z); \tag{22}$$

and equations (9), which express σ_{ZW^0}, $\sigma_{Z\bar{W}^0}$, and σ_w in terms of σ_{W^0}, $\sigma_{\bar{W}^0}$, and $\sigma_{\bar{W}^0W^0}$.

I have assumed that $M_{\bar{W}^0}$, N_{W^0}, $m_{\bar{W}^0}$, n_{W^0}, and θ (hence γ, M_Z, N_Z, m_Z, and n_Z) are observed for each group. Equations (9), (21), and (22) then may be solved simultaneously for the variances $\sigma^2_{W^0}$, $\sigma^2_{\bar{W}^0}$, and σ^2_w and the covariances $\sigma_{\bar{W}^0W^0}$, $\sigma_{Z\bar{W}^0}$, and σ_{ZW^0} for the group. The covariances σ_{ZW^0} and $\sigma_{Z\bar{W}^0}$, together with the data on $M_{\bar{W}^0}$, N_{W^0}, θ, and M_Z, then may be plugged into equations (19) and (12) to provide the estimates M_w and N_w of the net return to migration for migrants and nonmigrants, respectively, and the overall group means μ_{W^0} and $\mu_{\bar{W}^0}$. Thus, the procedure provides a way to estimate all of the underlying parameters μ_{W^0}, $\mu_{\bar{W}^0}$, σ_{W^0}, $\sigma_{\bar{W}^0}$, and $\sigma_{\bar{W}^0W^0}$ of the joint distribution of W^0 and $\bar{W}^0$ for the group from data on the mean and variance of destination net wages among migrants, the corresponding mean and variance of origin wages among nonmigrants, and the migration fraction θ. These estimates in turn may be used to estimate the selectivity biases in $M_{\bar{W}^0} - N_{W^0}$ as an estimate of M_w and N_w.

Progressivity and Investment Yields under U.S. Income Taxation

---

Martin J. Bailey

University of Maryland

Because tax deductions and tax credits associated with particular investments are worth more in equivalent pretax income to a high-bracket taxpayer than to a low-bracket one, tax-privileged investments find their way into the hands of high-bracket taxpayers. Their competition with each other and with the next lower brackets drive down pretax rates of return on such investments. As a result, although the tax system is less progressive than the rate structure alone makes it appear, it is more progressive than it would appear if we look only at net economic incomes and taxes. Tax privilege (exemption) relates qualitatively in a simple way to an investment's associated deductions and exclusions from income.

1. Introduction

It is well known, both among economists and among the lay public, that the U.S. tax system as a whole and the federal personal income tax in particular are less progressive in fact than the rate schedules make them seem. It is also well known among economists that, as Adam Smith first remarked, the tax system can induce people to misallocate resources (1937, p. 778). Much less understood is the interaction between these two phenomena, and the limits to them set by economic parameters. This paper sets out an analytical framework that explains and interprets the interaction, and that provides an improved theoretical basis for the results of studies of the incidence of U.S. income taxes. It shows, among other things, that the U.S. income tax must be proportional or progressive at the

---

I am indebted to Oswald Brownlee, Emil Sunley, Seymour Fiekowski, John Hall, and my other former Treasury colleagues for improving my perspective on income tax issues, and for constructive suggestions. Referees were also helpful. The errors that remain are mine.

[*Journal of Political Economy*, 1974, vol. 82, no. 6]

© 1974 by The University of Chicago. All rights reserved.

higher-income levels; any apparent drop-off in the effective rate of tax for very high incomes would be misleading because of sacrifices of pretax income to maximize income after tax. (Such sacrifices are part of the burden of a tax when we address questions of shifting and ultimate incidence.) It also shows that mild progressivity at high income levels is what we should expect on the basis of a check of the theory against rough benchmark data on income flows.

The complex web of tax provisions in the U.S. Internal Revenue Code results from a series of compromises among basic objectives and specific proposals. On any one topic, such as the deduction for travel expense or the child-care allowance, a comparatively generous proposal with a simple formula cannot command the votes to pass; and the still simpler proposal to do nothing also lacks the votes. The resulting formula that comes out of the tax-writing process is almost always more complex by far than either of the extremes. Of the compromises of this type, one of the most serious in its consequences and in its complexity is the covert bargain between those congressmen (and special interests) who want a highly progressive tax system and those who don't. Those who want it easily command the votes to enact the very progressive-looking rate schedule, but the lack the votes to make it stick. Those who oppose steep progressivity cannot prevent the enactment of the schedule, but they have ample support to riddle it with deductions, exclusions, and credits, worth more per dollar to wealthy people than to middle- and lower-income people. The effect of these deductions, exclusions, and credits is to moderate or eliminate the progressivity of the U.S. personal income tax above a certain income level—and to induce gross misallocation of resources. (For an earlier similar appraisal, see Simons 1938, p. 219.)

Early pathbreaking studies by Musgrave provided estimates of the overall incidence of the U.S. federal, state, and local tax systems combined, by income groups in the low and middle range, lumping together everyone over a low cutoff around $10,000 (Musgrave 1951, p. 26; 1955, p. 96.) Later, Musgrave estimated the incidence of the federal personal income tax in full detail for upper as well as lower income groups (1959*a*, p. 2223). A number of unresolved issues remained, some of which Musgrave himself pointed out. For overall incidence, the burden of the corporation income tax (and the proper distribution of corporate pretax earnings) was in doubt, as were the burden of the property tax and that of employer contributions to social security. Moreover, Musgrave found no way to account for accrued but unrealized capital gains for noncash income, or for reductions of income associated with burden shifting. His accounting for the use of privileged tax deductions to misstate income had to be incomplete because of data problems. (His final study took account of percentage depletion but not of several other special deductions and exclusions from income [1959*a*.]) These problems have been largely remedied, first by a series of papers on the microeconomic and general-equilibrium foundations of incidence analysis as applied to the corporation income tax (Harberger 1962; Krzyzaniak and Musgrave 1963; Krzyzaniak 1966; Cragg, Harberger, and Mieszkowski 1967) and more recently by a thorough study of tax and census data by Okner and Pechman (1973).

Okner and Pechman show that the tax system takes from about 20 to 25 percent of the incomes of all families except a few percent of families at each end of the income scale. For the small group of families with incomes above $100,000 per year, the effective burden of taxes ranges from 40 to 50 percent if one uses economic analysis to assess the burden, or could be as low as 30 percent if one discards such analysis in favor of ad hoc allocation. The taxes that determine the degree of progressivity for wealthy people are the personal and corporation income taxes.

2. Competition and After-Tax Rates of Return by Income Tax Bracket

Competition tends to equalize rates of return before taxes in a neutral-tax world; it equalizes them after taxes, among people in a given rate bracket, whether the world is neutral or not. Tax-favored activities come to equilibrium at lower pretax rates of return than normally taxed activities, so that after-tax returns tend to be equalized for taxpayers investing in both types of activities.[1]

If all sources of income were taxed alike, so that there was no way for the higher-income person to avoid or moderate the impact of higher rate brackets (apart from the labor-leisure choice and the consumption-savings choice), we would observe the same pretax yield on all investments and the simple after-tax pattern shown in figure 1. The yield before tax is r, and the after-tax yield i equals $r(1 - t)$; i declines linearly with the personal income tax rate, which ranges from 0 to 0.7 at the margin.

Tax-favored investments, with pretax returns driven down by competition, will look different to different taxpayers according to their marginal rate brackets. For example, a tax-exempt long-term bond yielding 5 percent is a poor investment for a person in the 14 percent tax bracket compared to a taxable Treasury bond yielding 7 percent. However, to a person in the 50 percent tax bracket, the tax-exempt bond is the better investment of the two. Likewise, any investment with favorable tax treatment, in equilibrium, will be more attractive to higher-bracket people than to lower-bracket ones. Moreover, the higher the tax bracket the greater is the relative advantage of the favorable tax treatment; and the higher the tax bracket, the more the pretax rate of return can be reduced

---

[1] For evidence in aggregated data, see Christensen (1971, p. 582).

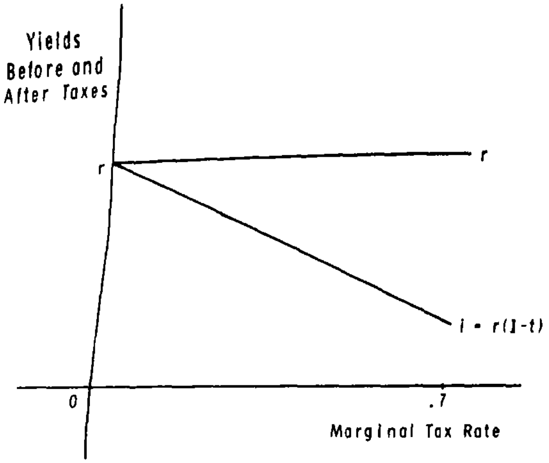


Fig. 1

before the advantage of this tax treatment is offset. The tax-exempt bond just mentioned is barely superior to the taxable bond for the person in the 30 percent bracket, whereas its "after-tax" yield is over twice that of the taxable bond (namely, 2.1 percent) for the person in the 70 percent bracket. If the yield of the tax-exempt bond declined to 3.5 percent, it would become a poor investment for the person in the 30 percent bracket, and indifferent to the person in the 50 percent bracket, while it would still look attractive to the person in the 70 percent bracket. He would prefer it until its yield dropped to or below 2.1 percent. The degree of these effects varies directly with the degree of tax exemption.

Figure 2 illustrates these points. As in figure 1, r and i represent the yields before and after tax on a fully taxable investment; r' and i' are the yields before and after tax on an investment half of whose income is taxable and half exempt; r'' is the yield on a completely tax-exempt investment. For persons in marginal rate brackets below t_A, given by the crossing of i and i' at A, the best available yield is i, from the fully taxable investment. Between t_A and t_B, the latter given by the point at which i' crosses r'', the half-taxable investment is the best of the three; persons in this range of marginal tax rates would prefer it to either the fully taxable or the tax-exempt investment. However, for the person in the highest marginal rate brackets, from t_B to 0.7, the tax-exempt is the most attractive of the three investments. Thus, as all types of taxpayers bid in the same markets for these different types of assets, yields on the most completely exempt investments are bid down until the amount available (and worth undertaking at these yields) is exactly absorbed by the highest-bracket people. Yields on partially taxed investments are bid down to a

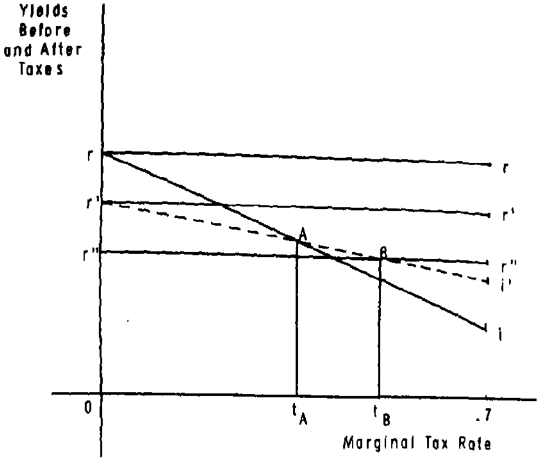


FIG. 2.—Fully taxed, semitaxed, and exempt investments

lesser extent, to the point at which the middle-bracket people to whom they are the best after tax are just able to absorb them. The most highly tax-exempt investments will settle into the portfolios of the highest-bracket taxpayers, the next most highly tax-exempt investments will settle into the portfolios of the next lower group of taxpayers, and so on down.

Moreover, there are some investments whose incomes are wholly tax exempt, and which in addition generate allowable excess deductions against other income. In such cases, the after-tax rate of return is greater than the pretax rate of return; we may characterize such investments as negatively taxed. (I use this term in preference to "subsidized" because some congressional and Treasury documents use the latter to cover both semitaxed and negatively taxed investments.) Figure 3 illustrates how the advantage of such a negative tax rises with marginal tax rate; the before-tax yield is r^*, and the after-tax yield is i^*, which rises in proportion to the marginal t times the average rate of flow of excess deductions. Such investments will find their way into the portfolios of investors who also hold tax-exempt investments.

In the following extreme example, a negative tax converts a negative social yield to a positive after-tax private yield. With ordinary, nonprivileged tax treatment, someone who buys woodland and clears it for resale as farmland would have to capitalize his clearing costs and deduct them for tax purposes only upon the sale of the land, to determine the profit from the sale. This profit would then be taxable as ordinary income. Under competition, we assume that the turnaround time for this operation is 1 year, and that with normal tax treatment the operator would have to

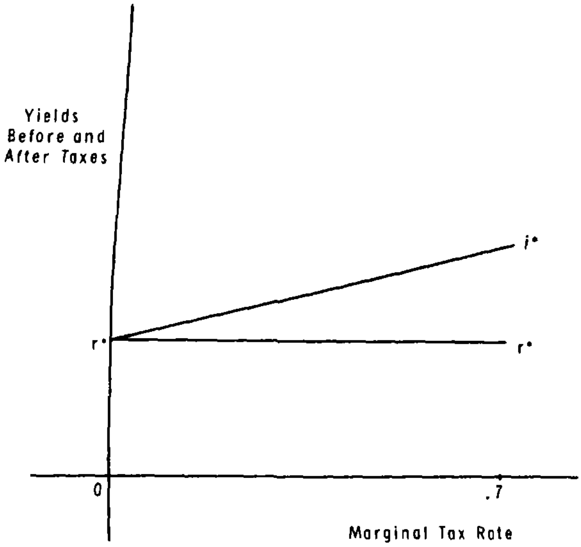


FIG. 3.—Tax-subsidized investment (negatively taxed)

TABLE 1

"LAND CLEARING"
CASH FLOWS

	Land Purchase	Clearing Cost	Tax Offset	Net Cost	Net Sale Price	Taxable Profit	Tax	Net Profit
Fully taxable . . .	−50	−50	0	−100	+120	+20	−14	+6
Tax subsidized . .	−50	−50	+35	−65	+80	+15	−10.5	+4.5

earn 6 percent after tax to justify the investment. The first row of table 1 shows such a case, where the woodland costs 50 and the clearing cost is 50. With no deductible expenses, the operator invests 100 of his own money (shown as "net cost"); for an adequate return, he must get at least 120 in net sale proceeds the next year. If so, his taxable income from the transaction is 20, on which he pays the full marginal rate of 0.7, for a tax of 14 and a net profit of 6.

Now suppose that the clearing cost can be written off against other income in the first year, and that the profit on sale—the excess of sale price over the "cost basis" consisting only of the land purchase price—is taxed as a long-term capital gain. In the first year the operator gets a tax deduction of 50, for a tax offset of 35 (at the 0.7 marginal rate), so that his own investment is only 65. In the second year he can afford to sell for as little as 80, for a social loss of 20 percent of the total investment.

TABLE 2

RECOVERY OF INVESTMENT BY CONTINUING DRILLING FUNDS AS OF 1972

Drilling Fund Start-Up Year	Total Investment* ($ Millions) (1)	Capital Recovered as Income ($ Millions) (2)	Fraction Recovered (2) ÷ (1)
1964	35.3	23.9	0.676
1965	40.1	24.2	0.605
1966	39.5	20.6	0.521
1967	105.9	66.6	0.629
1968	135.0	47.6	0.353
1969	263.5	58.8	0.223
1970	253.3	47.6	0.188
1971	241.0	12.0	0.050

SOURCE.—Unpublished Treasury data.
* Includes borrowed capital.

His sale price exceeds his cost basis by 30, of which only half is includable in his taxable income. (We ignore the alternative tax on capital gains, which would make the picture even worse.) His tax is 0.7 times 15, or 10.5, leaving a net profit of 80 − 65 − 10.5 = 4.5. On his own investment of 65, his rate of return is 4.5/65 = .0693, or almost 7 percent. Extreme cases of negative tax like this do occur from time to time.[2] Note, however, that competition drives down the price margin between cleared and uncleared land until the net rate of return to this investor is about the same as he could get on taxed investments.

Treasury data, previously unpublished, on syndicated oil "drilling fund" operations illustrate this last point. The figures shown in table 2 for each year are the total investment and income through early 1972 of drilling funds started in the indicated year, cumulated without interest; the funds included in the sample are those that had registered public offerings in the first nine months of 1972 and had made all their past payouts in cash. The oldest funds, started from 1964 to 1967, had got back about 60 percent of their total investment, plus unstated depreciation allowances on capitalized costs; detailed data, not shown, show that their earnings were falling off by 1972, as must be expected with old oil wells. Depreciation adds roughly another 20 percent to capital recovery (Agria 1969, p. 97). However, it is unlikely that the group of investors represented in this table will ever get a reasonable return on their total investment, even after allowing for a small degree of negative taxation (by the criteria of the next section, below).

The data in table 2 have opposite biases that are hard to estimate. Drilling funds that started after 1963 but failed before 1972 are not shown; neither are those that were still in business but made no public offerings in

---

[2] I am indebted to Finis Welch for the main outline of this example.

TABLE 3

OPTIMAL $10 MILLION PORTFOLIO

Investment	Tax Exempt: $7 Million	Semitaxed: $3 Million	Total: $10 Million
Gross income	350,000	200,000	550,000
"Adjusted gross income"	0	100,000	100,000
Deductions, etc.	...	...	50,000
Taxable income	...	...	50,000
Tax	...	...	17,060
Net income after tax	...	...	532,940

January–September 1972. Also excluded are some that had redeemed some of their investors' claims with corporate shares; these would on balance be more profitable than average. Finally, the best drilling opportunities, among others, are likely to be exploited by some other means of financing than a drilling fund; in fact, these funds accounted for only a small part, less than 20 percent, of noncorporate wildcatting. However, these data do represent a large part of a standard type of organized tax shelter investment, and so are pertinent to the present argument. To the extent that these numbers are representative, they support the argument, if anything a little too well. The main profit in these operations is for their promoters.

We should further note that the profitable strategies of persons holding investments of this type are limited by the necessity of having otherwise taxable income against which to deduct the allowable deductions from a privileged investment. Without some taxable income, a taxpayer would not be in the 70 percent bracket. Therefore it would never pay to have a portfolio consisting only of a mixture of negatively taxed investments and tax-exempt bonds, or indeed a portfolio consisting only of tax-exempt bonds, unless salary, fee income, and the like provided enough taxable income to justify such holdings

For example, consider a person whose income-earning assets are worth $10 million and who has no salary or other personal-service income. Suppose also that a tax-exempt investment he wants in his portfolio yields 5 percent annually, while a semitaxed investment, half of whose income is tax exempt, yields 6.67 percent. These two investments have the same net return after taxes to a person in the 50 percent tax bracket. If this person puts all his $10 million into the tax-exempt investment, his income will be $500,000 per year. Were he to make the mistake of investing entirely in the semitaxed investment, if his allowable deductions were of the order of $50,000 his net income after tax would be about $425,000. However, he should buy some of this investment, enough to generate a taxable income, after deductions, of around $50,000. If he buys the portfolio shown in table 3, he has gross income of $350,000 from his tax-exempts and \$200,000 from his semitaxed investment, giving an adjusted gross income on his tax return of \$100,000. His taxable income, after deductions, of \$50,000 gets him into the 50 percent bracket, justifying his holding of tax-exempts. By absorbing his deductions and the lower bracket rates, his \$3 million of semitaxed investment clears more for him than would an extra \$3 million of tax-exempt investment: his net income after tax winds up \$32,940 higher than it would be if he invested only in tax-exempts. Thus, any wealthy person who "shelters" his income so much as to have no taxable income has either mismanaged his portfolio or has been unlucky. (A conceivable exception to this rule is a person whose specialized personal talents combine with all the capital he can profitably own and borrow in a negatively taxed activity. This activity may provide his best available opportunity, even though it generates "wasted" tax deductions; some land developers are said to be typically in this position.)

In summary, competition and portfolio optimization by taxpayers will lead to a pattern of declining before-tax rates of return as a function of marginal rate brackets. After-tax rates of return decline also for higher rate brackets, and become flat to the extent that tax-exempt investments are available (whose after-tax yields are independent of tax bracket, because tax is zero). Negatively taxed investments have no effect on this picture because to get the negative tax the investor must combine them with partly taxable investment, to generate taxable income not less than zero for the combination. The combination is then a tax-exempt investment from the standpoint of incidence, although the harmful resource allocation effects just noted for the negatively taxed cases still hold. (In this discussion, I disregard taxpayers with large personal-service incomes.) Figure 4 represents these conclusions, showing the pretax yield r' declining as a function of t, and the after-tax yield i also declining but flattening out at the highest values of t.

The corporation income tax, being unrelated to the tax bracket of the shareholder, is virtually a proportional tax on corporate-source income. I show the pretax yield on corporate investment as r_c, higher than r because it pays corporation income tax in addition to the personal income tax of the shareholder. Shareholders invest on the basis of expected returns net of corporation income tax, and, as noted earlier, the data show that these net returns are equalized. Because corporations differ in the extent to which the yield on their shares can be expected to accrue as capital gains, high-bracket taxpayers tend to hold growth stocks, whereas low-bracket tax-payers tend to hold income stocks. All shares that generate capital gains give semitaxed income to their shareholders, and so should in equilibrium be held by high-bracket taxpayers (see Bailey 1969, pp. 35–37). After discounting for risk, therefore, the yields after corporation income tax but before personal income tax should match the r' yield curve. Then r_c is obtained as a fixed multiple of r' at each tax bracket, as

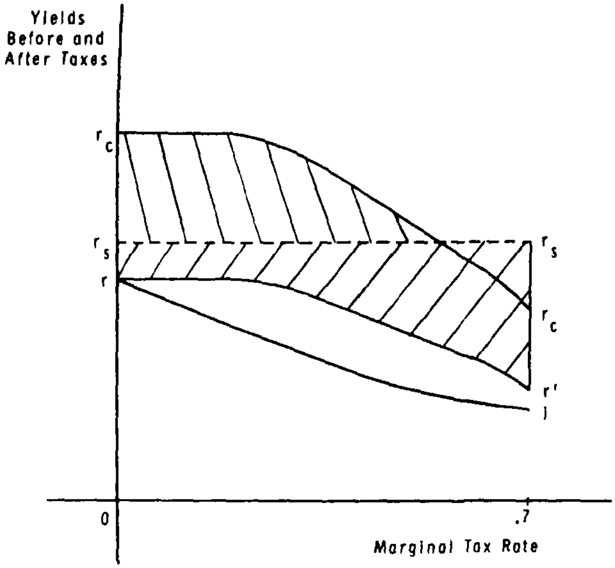


Fig. 4

shown; if the effective corporation income tax rate is 40 percent, the multiple is 1.67.

The extra income and tax on corporate investment are imputable in part to persons other than shareholders, such as the owners of other capital. In a tax-neutral world there would be some social return to investment, the same for every type, equal to r_s. The excess of r_c over r_s, all or more than all of which is offset by the corporation income tax, would be excluded from shareholder income in a full analysis of shifting and incidence, and this portion of the corporation income tax (that offsets the excess) would be attributed as a burden to other people. Likewise, the entire difference between r_s and i represents the true burden of corporate and personal income taxes on the owners of capital, arrayed by tax bracket. (In this discussion I assume the correctness of Harberger [1962], which has yet to be successfully controverted.) Moreover, for higher-bracket people the true burden of the personal income tax alone is the difference between r and i. In those studies that look only at the difference between r' and i, that is, at the tax paid directly by taxpayers as a proportion of their money incomes, the burden on higher-bracket taxpayers is understated.

This conclusion implies, incidentally, that it is misleading to assume that there is no shifting of the personal income tax where factor supplies are inelastic. To the extent that the i-curve flattens out instead of having the straight linear decline of figure 1, high-bracket taxpayers simply avoid the personal income tax; but that is not the whole story. To the

TABLE 4

CASH FLOWS FOR 1-YEAR INVESTMENT

	First Year (Outlay)	Tax Offset	First Year (Net Outlay)	Second-Year Receipts	Tax	Net Profit	Profit (%)
Standard tax treatment	−100	0	−100	+110	−7	+3	+3
Immediate write-off........	−100	+70	−30	+110	−77	+3	+10

extent that r' drops below r, high-bracket taxpayers have accepted a tax burden, "shifted" by their portfolio choices, that exceeds their apparent burden by the difference between r' and r.

3. Deductions and the Degree of Tax Privilege

In the preceding section we noted that overgenerous deductions associated with some investments confer partial tax exemption or a negative tax on them, but we gave no indication of the precise connection between deductions and tax exemption. In this section we show this connection, and the boundary line of effective tax exemption.

It has long been known that when an investment can be written off completely at the time it is made, it is effectively tax-exempt—the after-tax rate of return equals the before-tax rate of return (Musgrave 1959*b*, pp. 343–44. Rolph [1954, p. 279] attributes the idea to Davenport.) For example, suppose that a 1-year investment yielding 10 percent can be "expensed" immediately, making the full positive cash flow, the return of principal and interest in the second year, taxable. The net investment in the first year, after tax offset, is only 30 for an investor in the 0.7 bracket, so that his after-tax net profit of 3 in the second year is 10 percent of his investment, compared to 3 percent if he has to use inventory accounting and deduct the investment outlay from the sale proceeds in the second year (see table 4). The same result follows for an investment of any life; indeed I shall generalize this proposition into a criterion covering all types of special tax treatment.

Consider as a "deduction" D any of the following tax provisions affecting a particular investment:

1. Depreciation allowances for tax purposes.
2. Other exclusions from taxable income associated with an investment, such as percentage depletion allowances for minerals, and the excluded half of long-term capital gain to the extent that such a gain will be part of the outcome of the investment.
3. The pretax income equivalent to a tax credit that is contingent on

this investment, such as the investment tax credit. That is, if t is the marginal tax rate of the investor and C is the credit, $D = C/(1 - t)$.

4. If the investor expects to be in a lower marginal rate bracket when he receives taxable cash flows from the investment than he is at the time of claiming deductions against other income, that reduction of taxable income that would produce the same tax reduction on these cash flows as he gets by virtue of the lower marginal rate. That is, if his rate now is t and it will be t' in year k when he expects a taxable cash flow R_k from the investment, we define the implied deduction D_k by the relationship $t(R_k - D_k) = t'R_k$, or $D_k = [(t - t')R_k]/t$. In this formula R_k is only the taxable portion of the cash flow of year k, net of any deductions of the first three types that hinge on this investment. Such other deductions are then added to the D_k just shown to get the total implied deduction for year k.

With deductions so defined, we have the following proposition: *The after-tax rate of return on an investment equals the before-tax rate of return if and only if the present value of all its deductions equals the present value of the investment outlay.* We prove this proposition for the case in which the investment outlay all comes immediately, having a present value equal to the undiscounted outlay, followed by nonnegative net receipts before tax and after tax. (Extension to other cases is straightforward.)

Let the before-tax rate of return r be the solution to

$$\sum_k \frac{R_k}{(1 + r)^k} = I, \tag{1}$$

where I is the initial investment outlay and where the later pretax cash flows R_k are all nonnegative, with some positive, which implies the uniqueness of r (Hirschleifer 1970, pp. 77–79). Let the after-tax cash flows be Z_k, defined by

$$Z_k = R_k - t(R_k - D_k) > 0, \tag{2}$$

also nonnegative, giving uniqueness of the implied rate of return i:

$$\sum_k \frac{Z_k}{(1 + i)^k} = I. \tag{3}$$

The first part of the italicized proposition above says that if the present value P of deductions equals the investment,

$$P = \sum_k \frac{D_k}{(1 + i)^k} = I, \tag{4}$$

then $r = i$. Substituting (3) and (2) into (4), we have

$$\sum \frac{D_k}{(1 + i)^k} = \sum \frac{Z_k}{(1 + i)^k} = \sum \frac{R_k(1 - t) + tD_k}{(1 + i)^k}.$$

Then, rearranging,

$$\sum \frac{R_k(1 - t)}{(1 + i)^k} = \sum \frac{D_k(1 - t)}{(1 + i)^k}$$

or

$$\sum \frac{R_k}{(1 + i)^k} = \sum \frac{D_k}{(1 + i)^k}; \tag{5}$$

and by (1) and (4) we can replace the right-hand side to give

$$\sum \frac{R_k}{(1 + i)^k} = \sum \frac{R_k}{(1 + r)^k},$$

which implies $r = i$ as proposed, because r is unique.

The other half of the proposition, "only if," means that $r = i$ implies $P = I$. We have

$$\begin{aligned} \sum \frac{R_k}{(1 + r)^k} &= \sum \frac{R_k}{(1 + i)^k} \\ &= \sum \frac{Z_k}{(1 + i)^k} \\ &= \sum \frac{R_k(1 - t) + tD_k}{(1 + i)^k}. \end{aligned}$$

Using the right-hand sides of the first and third of these equalities, and collecting terms, we find

$$\sum \frac{D_k}{(1 + i)^k} = \sum \frac{R_k}{(1 + i)^k} = I, \tag{6}$$

which completes the proof. Furthermore, by evaluating derivatives, we can show that $i \lesseqgtr r$ according as $P \lesseqgtr I$, and conversely.

Thus we have a general characterization for investments that are negatively taxed: the present value of the deductions, discounted using i, exceeds the investment. This present value also exceeds the present value of the future pretax cash flows, which should not surprise us. The case in which the only deduction is the immediate write-off of the investment is a special case of the above result; also, a trivial special case is that in which the investment is nondeductible but in which future cash flows are tax exempt, so that the deduction always equals the cash flow. We see further that any investment that can be immediately written off, and whose future cash flows are partly tax exempt (as happens with capital gains treatment of cleared land in the example of table 1, and happens approximately also with oil wells), is negatively taxed.

We can usefully carry the inequalities further. From equations (5) and (6) and their derivations, we have

$$\sum \frac{D_k}{(1+i)^k} \lesseqgtr \sum \frac{R_k}{(1+i)^k} \text{ according as } i \lesseqgtr r, \tag{7}$$

and conversely. Now, for any given investment project, I, R_k, D_k, and r are fixed and independent of tax bracket. Thus equations (2) and (3) give an implicit function between i and t:

$$I = \sum \frac{R_k(1-t) + tD_k}{(1+i)^k}.$$

Taking differentials and rearranging, we obtain

$$\frac{\delta i}{\delta t} = \frac{\sum [(D_k - R_k)/(1+i)^k]}{\sum [kZ_k/(1+i)^{k+1}]}.$$

The assumption that all $Z_k > 0$ means that the denominator of this derivative is positive. By (7), we have that its numerator is positive, zero, or negative according as i is greater than, equal to, or less than r.

Thus we have established that negatively taxed investments have higher after-tax rates of return for high bracket taxpayers than for low-bracket taxpayers, whereas partially or fully taxable investments have lower after-tax rates of return for high-bracket taxpayers than for low-bracket taxpayers, and conversely. A tax-exempt investment has the same after-tax rate of return for everyone—regardless of the time profile of the associated deductions. (This conclusion and others depending on equal D_k for different investors have to be qualified where the investor expects a lower tax bracket for his later R_k than for his early deductions.) These results justify the slopes of the i-curves in figures 2 and 3.

4. Progressivity in Equilibrium

In Section 2, we have a general picture of the degree of progressivity implied in the combined provisions of the U.S. Tax Code, but the theory is too general to tell us how much less progressivity we should expect than that implied by the tax schedule. The extent to which tax privilege flattens out the after-tax yield for higher-bracket taxpayers, the i-curve in figure 4, depends on the volume of available privileged investments at acceptable yields, relative to the volume of investment financed by higher-bracket taxpayers. Conceivably, the volume of such investment might be so small that the highest-bracket taxpayers would absorb them all without their having any effect on progressivity; or their volume might be so great that there would be no taxpayers in the statutory highest bracket. Various sources, including the Okner-Pechman study (1973), indicate

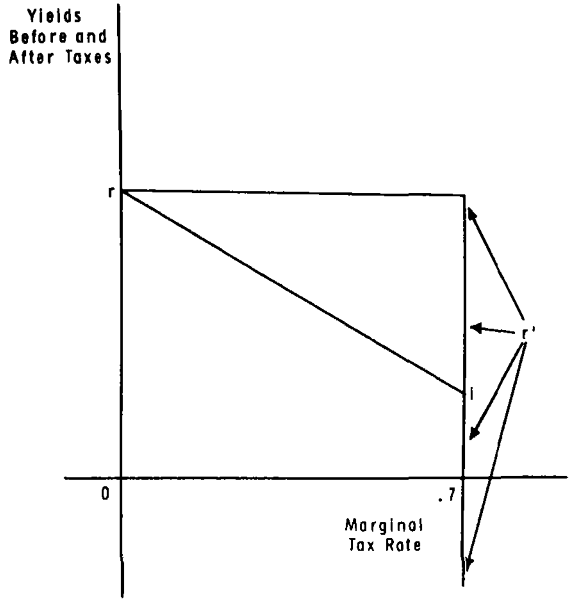


FIG. 5.—Case *A*

that we are between these extremes. We can address this question directly.

Consider the following three cases.

Case *A*: All available negatively taxed, tax-exempt, and semitaxed investment goes to persons in the 0.7 bracket, who have money left over to invest in some fully taxed investment.

Case *B*: All available negatively taxed and tax-exempt investment goes to persons in the 0.7 bracket, who must also invest in some semitaxed investment in addition to that necessary to offset the deductions coming from the negatively taxed investment. Some semitaxed investment goes to taxpayers in brackets below 0.7.

Case *C*: The volume of negatively taxed investment is so great that some of it goes to taxpayers in brackets below 0.7.

If Case *A* were true, the relationship between i and r at the margin would be that shown in figure 5. The after-tax rate of return i would be linear in t, as in figure 1, and the investment yielding r before tax would be financed by taxpayers in all tax brackets. Some of the investment financed by taxpayers in the 0.7 bracket would be the privileged investment with lower pretax yields, but all their investment would have the after-tax yield $i = (0.3)r$ given by the end of the linear i curve.

If Case *C* were true, the relationship among r, r', and i would be that

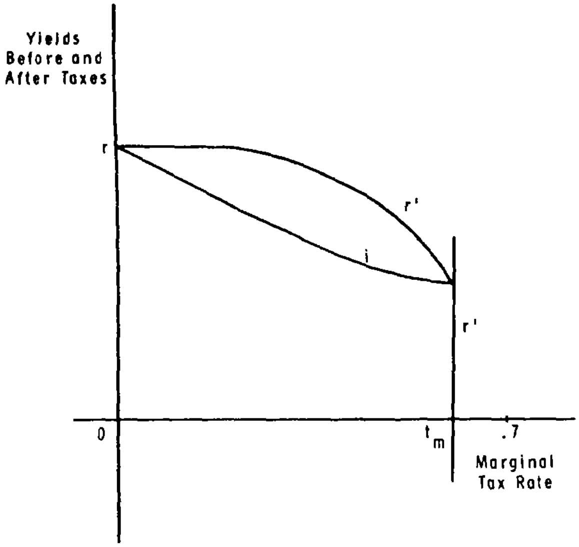


FIG. 6.—Case *C*

shown in figure 6. Taxpayers potentially in the 0.7 bracket would finance so much tax-privileged investment that they would all stay in a lower maximum tax bracket t_m. The declining r' curve would touch the i curve at t_m, so that all taxpayers at that maximum rate would add only tax-exempt combinations of investment to their portfolios. (Personal service income and income from old assets would provide taxable income to push them to t_m.) We would observe no one in the 0.7 bracket, except insofar as risk and other factors led some to accept less than the maximum yield after taxes.

If Case *B* is true, we expect the relationship among r, r', and i shown in figure 7. The after-tax yield i declines all the way to the 0.7 bracket, with some taxpayers remaining in this bracket. The pretax yield r' remains above i for all taxpayers below 0.7, and also for some at 0.7. However, at 0.7 their pretax yields range down into the negatively taxed values. It seems clear that this case is the most plausible of the three, as it is consistent with Okner-Pechman and with earlier studies.

If all taxpayers were alike, we could rule out Case *C* because there are large amounts of income taxed at the top marginal rate. About $1.5 billion belonged to taxpayers in this rate bracket in 1970; this figure is about one-fifth of the aggregate incomes of persons having $200,000 per year or more, which is roughly the group capable of getting into the 70 percent bracket (U.S. Treasury Department 1970).

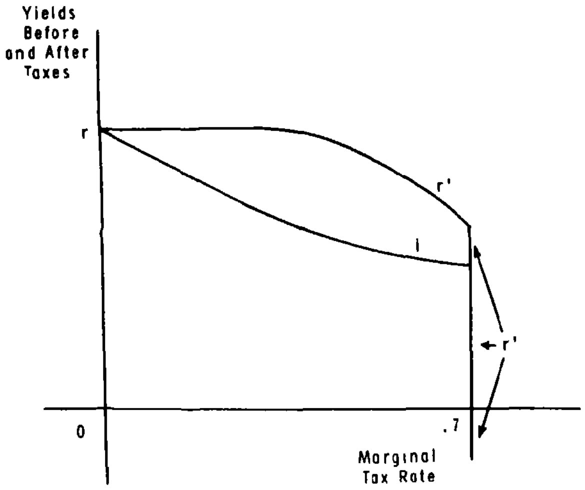


FIG. 7.—Case *B*

Although data on the volume of negatively taxed and partially tax-exempt investment are unavailable, we know that all forms of residential housing are at least partly tax exempt. Moreover, corporate shares, especially growth stocks, are partly tax exempt from the viewpoint of the shareholder because he expects capital gains tax treatment of a large part of their income. These possibilities include tens of billions of net investment ($54 billion in residential housing alone in 1972 [see U.S. Office of the President 1974, table C-13]), far more than could be financed by the top-bracket group. Thus we can even more surely rule out Case *A*. We are left with Case *B* as the one most consistent with available evidence.

However, if we classify taxpayers by their degree of risk aversion, we might reach a mixed verdict. Experienced tax lawyers know of many taxpayers with incomes in the range from $100,000 to $200,000, that is, with marginal rates ranging from 58 percent to 69 percent, who invest in tax shelters. This observation seems compatible only with Case *C*. As it happens, negatively taxed "tax shelter" investments range from risky to disastrous, and so go to persons with less than average risk aversion. Such persons are evidently scarce enough to put them as a group into Case *C*. In contrast, the relatively risk-averse parts of the wealthy population go into Case *B*, deriving most of their income from capital gains unaided by any other significant tax privilege, because acceptably low-risk tax shelter investments are unavailable in large enough quantities to be a significant part of their portfolios.

5. Conclusion

The overall burden of the personal income tax is only mildly progressive for incomes in the higher marginal rate brackets, because of the moderating effect of partial tax exemption for a large fraction of total investment. Apparent horizontal inequities as a rule shake out in competitive resource allocation and translate into misuse of resources. Taxpayers respond to tax inducements by overinvesting in the favored industries, and although such taxpayers appear to pay less than their "fair shares" of tax, their tax avoidance accomplishes no more than to make effective personal income tax burdens approximately proportional to the higher income ranges. This limited accomplishment is due to the driving down of pretax rates of return r' by competition. The lowered rates of return reduce real GNP, as a penalty for the covert political bargain that produced this complex system.

References

Agria, Susan R. "Special Tax Treatment of Mineral Industries." In *The Taxation of Income from Capital*, edited by Arnold C. Harberger and Martin J. Bailey. Washington: Brookings Inst., 1969.

Bailey, Martin J. "Capital Gains and Income Taxation," in *The Taxation of Income from Capital*, edited by Arnold C. Harberger and Martin J. Bailey. Washington: Brookings Inst., 1969.

Christensen, L. R. "Entrepreneurial Income: How Does It Measure Up?" *A.E.R.*, vol. 61, no. 4 (September 1971).

Cragg, John C.; Harberger, Arnold C.; and Mieszkowski, Peter. "Empirical Evidence on the Incidence of the Corporation Income Tax." *J.P.E.*, vol. 75, no. 6 (December 1967).

Harberger, Arnold C. "The Incidence of the Corporation Income Tax." *J.P.E.*, vol. 70, no. 3 (June 1962).

Hirschleifer, J. *Investment, Interest, and Capital.* Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Krzyzaniak, Marian, ed. *Effects of the Corporation Income Tax.* Detroit: Wayne State Univ. Press, 1966.

Krzyzaniak, Marian, and Musgrave, Richard A. *The Shifting of the Corporation Income Tax.* Baltimore: Johns Hopkins Press, 1963.

Musgrave, Richard A. "The Distribution of Tax Payments by Income Groups." *Nat. Tax J.* vol. 4, no. 1 (March 1951).

———. "The Incidence of the Tax Structure and Its Effects on Consumption." In *Federal Tax Policy for Economic Growth and Stability, Papers Submitted by Panelists* . . . , Joint Committee on the Economic Report. Washington: Govern- Printing Office, 1955.

———. "How Progressive Is the Income Tax?" In *Tax Revision Compendium*, House Committee on Ways and Means. Vol. 3. Washington: Government Printing Office, 1959. (*a*)

———. *The Theory of Public Finance.* New York: McGraw-Hill, 1959. (*b*)

Okner, Benjamin A., and Pechman, Joseph A. "Who Paid the Taxes in 1966?" Paper delivered to the American Economic Association, New York, December 30, 1973. *American Econ. Assoc., Proc.* (in press).

Rolph, Earl R. *The Theory of Fiscal Economics.* Berkeley and Los Angeles: Univ. California Press, 1954.

Simons, Henry. *Personal Income Taxation.* Chicago: Univ. Chicago Press, 1938.

Smith, Adam. *The Wealth of Nations.* New York: Random, 1937.

U.S., Office of the President. *Economic Report of the President.* Washington: Government Printing Office, 1974.

U.S., Treasury Department, Internal Revenue Service. *Statistics of Income: Individual Income Tax Returns, 1970.* Washington: Government Printing Office, 1970.

Adam Smith on Profits—Paradox Lost and Regained

Nathan Rosenberg

Stanford University

This paper examines the following apparent paradox. Adam Smith's *Wealth of Nations* is universally regarded as a book that powerfully presented the social case for giving the businessman the maximum degree of freedom of action. And yet, although Smith unqualifiedly treats high wages as desirable, he treats high profits as undesirable. Why did so eloquent a spokesman for capitalism and laissez faire treat high profits with such unrelieved hostility? The interpretation of Smith presented in answering this question revolves primarily around the belief that easily earned high profits destroy the effectiveness with which the capitalist carries out his social role.

It might be best to begin with a candid admission. This paper is not the paper I had originally set out to write. Adam Smith's treatment of the business community in general and the entrepreneur in particular is a subject that has long fascinated me.[1] It seemed therefore to be both useful and rewarding to ferret out and to examine Smith's treatment of the role of the entrepreneur in the *Wealth of Nations*.[2] After all, in a book that has been regarded as the *locus classicus* of the laissez faire ideology for 200 years, a book that shook the world by recommending a maximum degree of freedom for business enterprise—in such a book surely the entrepreneur would play a major role. Although this originally seemed like a reasonable expectation, it was not fulfilled. As I should certainly have realized, Smith's analytical distinctions here were inevitably limited by the modest state of capitalist development itself and the relatively small degree of specialization of function that still prevailed in the middle of the eighteenth century. Indeed, Smith had made a significant contribution to

---

[1] See Rosenberg (1960).

[2] Smith (1937).

[*Journal of Political Economy*, 1974, vol. 82, no. 6]

© 1974 by The University of Chicago. All rights reserved.

analytical economics merely by his forceful recognition of profit on capital as constituting a separate and distinct income category.[3]

I have not, I am happy to report, returned from this brief excursion into intellectual history completely empty-handed. Instead of the story I had hoped to tell, I want to report upon a rather unexpected paradox that I encountered. I would like, moreover, to try to unravel this paradox. My determination to do so comes about not only because of a certain taste for intellectual history—although that would be justification enough—but also because the paradox is one that goes to the heart of Adam Smith's Weltanschauung.

Stated in somewhat oversimplified terms, my paradox lies in the fact that Adam Smith treated high wages as being unqualifiedly a good thing, and high profits as being unqualifiedly a bad thing. Why should so eloquent a spokesman for capitalism and laissez faire regard high profits with such a jaundiced eye? Indeed, he closes his chapter on profits with the following devastating barrage: "Our merchants and master-manufacturers complain much of the bad effects of high wages in raising the price and thereby lessening the sale of their goods both at home and abroad. They say nothing concerning the bad effects of high profits. They are silent with regard to the pernicious effects of their own gains. They complain only of those of other people."[4] So fervently did Smith believe these sentiments that the statement actually appears, substantially unchanged, in two different places in *The Wealth of Nations.*[5]

The question of whether high wages were desirable had such an obviously affirmative answer to Smith that he did not even undertake to justify it, but rather asserted it with a rhetorical flourish: "Is this improvement in the circumstances of the lower ranks of the people to be regarded as an advantage or as an inconveniency to the society? The answer seems at first sight abundantly plain. Servants, labourers and workmen of different kinds make up the far greater part of every great political society. But what improves the circumstances of the greater part can never be regarded as an inconveniency to the whole. No society can surely be flourishing and happy, of which the far greater part of the members are poor and miserable."[6]

---

[3] On this subject, see the perceptive article by Meek (1954). Meek states: "Many of Smith's predecessors had recognized, of course, that those who employed stock in mercantile pursuits generally received a net reward which was proportioned not to the effort, if any, which they had expended, but rather to the value of the stock employed. In Smith's new model it was recognized that net gains similar in this respect to mercantile profit were now also being earned on capital employed in other economic pursuits, such as agriculture and manufacture. But even more important, it was also recognized that the *origin* of these net gains was now very different from what it had formerly been. To Smith's predecessors, generally speaking, profit had appeared as 'profit upon alienation'—i.e., as the gain from buying things cheap and selling them dear. To Smith, on the other hand, profit began to appear as an income uniquely associated with the use of capital in the employment of wage-labor" (pp. 138–39).

[4] Smith (1937, p. 98).

[5] "Our merchants frequently complain of the high wages of British labour as the cause of their manufactures being undersold in foreign markets; but they are silent about the high profits of stock. They complain of the extravagant gain of other people; but they say nothing of their own. The high profits of British stock, however, may contribute towards raising the price of British manufactures in many cases as much, and in some perhaps more, than the high wages of British labour" (Smith 1937, pp. 565–66). This statement appeared in the first edition of the book whereas the one in the text above made its appearance in the second edition.

If Adam Smith's views here do not strike our egalitarian sensibilities as particularly startling, that is partly because of a drastic shift in attitudes which Smith himself played some role in bringing about. For earlier in the eighteenth century the dominant view was of the social utility of poverty. Smith's predecessors were very much exercised—indeed, some were absolutely obsessed—over the socially undesirable consequences of high or rising wages. It had been a firmly accepted part of the conventional wisdom that high wages would reduce effort, that the working class response to higher wages could be described—in the jargon of a later day—in the form of a backward-sloping labor-supply curve. The dominant view was well expressed by Arthur Young, that repository of conventional wisdom, who wrote in his *Farmer's Tour through the East of England* in 1771 that "every one but an idiot knows that the lower classes must be kept poor or they will never be industrious; I do not mean, that the poor of England are to be kept like the poor of France, but, the state of the country considered, they must (like all mankind) be in poverty or they will not work."[7]

Similarly, Sir William Temple, in his *Vindication of Commerce and the Arts*, says categorically of laborers that "the only way to make them temperate and industrious, is to lay them under the necessity of labouring all the time they can spare from meals and sleep, in order to procure the necessaries of life."[8] Such were the dominant views of Smith's time.[9]

In such a context, Smith's views were both enlightened and advanced. Moreover, they were novel in a respect that needs to be made quite explicit. Not only did Smith believe that high wages were intrinsically desirable because they improved the standard of living of the mass of the population, but he also believed—and here he clashed head-on with the prevailing view—that the working class supply of effort was positively sloped, that higher wages called forth greater effort and not less: "The liberal reward of labour . . . increases the industry of the common people. The wages of labour are the encouragement of industry, which, like every other human quality, improves in proportion to the encouragement it receives. A plentiful subsistence increases the bodily strength of the labourer, and the comfortable hope of bettering his condition, and of ending his days perhaps in ease and plenty, animates him to exert that strength to the utmost. Where wages are high, accordingly, we shall always find the workmen more active, diligent, and expeditious, than where they are low."[10] Although Smith concedes that higher wages are likely to induce some workers to reduce the number of hours worked, he is insistent that such workers constitute only a minority of the labor force. Indeed, Smith appears to be genuinely concerned over the opposite possibility, that a system of incentive wages will cause many workers to suffer the deleterious effects of overwork. In this respect he is the first economist of whom I am aware for whom this was a major concern.

---

[6] Smith (1937, pp. 78–79).

[7] Young (1771, 4: 361).

[8] Temple (1786, p. 534).

[9] For an excellent scholarly presentation of the mercantilists' attitude toward labor, see Furniss (1957). For a careful study of the transition from the old set of views to the later ones, see Coats (1958). Coats states: "Apart from a few isolated advocates of a 'high wage economy,' most British economists before 1750 regarded low wages as an essential precondition of the maintenance of a high volume of exports, although the plea that the British workman should enjoy a higher standard of living than that of his continental counterpart represented a tacit admission that successful competition in foreign markets did not require that home wage levels should be equal to or lower than foreign wage levels. By contrast, in the third quarter of the century there was growing support for the view that high wages and rising living standards were not merely compatible with, but were even a necessary concomitant of the prosperity of our domestic and exported manufactures" (p. 46).

> Some workmen, indeed, when they can earn in four days what will maintain them through the week, will be idle the other three. This, however, is by no means the case with the greater part. Workmen, on the contrary, when they are liberally paid by the piece, are very apt to over-work themselves, and to ruin their health and constitution in a few years. A carpenter in London, and in some other places, is not supposed to last in his utmost vigour above eight years. Something of the same kind happens, in many other trades, in which the workmen are paid by the piece, as they generally are in manufactures, and even in country labour, whenever wages are higher than ordinary.[11]

Where workers do, in fact, avail themselves of long intervals of leisure, Smith finds the cause, not in laziness or deficiency of character, but in deeply rooted physiological causes.

> Excessive application during four days of the week, is frequently the real cause of the idleness of the other three, so much and so loudly complained of. Great labour, either of mind or body, continued for several days together, is in most men naturally followed by a great desire of relaxation, which, if not restrained by force or by some strong necessity, is almost irresistible. It is the call of nature, which requires to be relieved by some

---

[10] Smith (1937, p. 81).

[11] Smith (1937, pp. 81–82).

> indulgence, sometimes of ease only, but sometimes too of dissipation and diversion. If it is not complied with, the consequences are often dangerous, and sometimes fatal, and such as almost always, sooner or later, bring on the peculiar infirmity of the trade. If masters would always listen to the dictates of reason and humanity, they have frequently occasion rather to moderate, than to animate the application of many of their workmen.[12]

When Smith turns from the examination of the economic behavior of the worker to that of the capitalist, his attitude shifts from that of compassion and understanding to one of compulsive and cantankerous criticism and suspicion. The long-term interests of capitalists, to begin with, do not coincide with those of society: "The rate of profit does not, like rent and wages, rise with the prosperity, and fall with the declension, of the society. On the contrary, it is naturally low in rich, and high in poor countries, and it is always highest in the countries which are going fastest to ruin. The interest of this third order [i.e., capitalists], therefore, has not the same connexion with the general interest of the society as that of the other two [i.e., landlord and worker]."[13] As a result, capitalists as a class are simply not to be trusted: "The proposal of any new law or regulation of commerce which comes from this order, ought always to be listened to with great precaution, and ought never to be adopted till after having been long and carefully examined, not only with the most scrupulous, but with the most suspicious attention. It comes from an order of men, whose interest is never exactly the same with that of the public, who have generally an interest to deceive and even to oppress the public and who accordingly have, upon many occasions, both deceived and oppressed it."[14]

A businessman who had been taught to regard Adam Smith as a capitalist apologist might well be excused for wondering what sort of strange capitalist apologetics this is, and, if this is what we are likely to get from our friends, just what may we expect from our enemies. Part of the answer may be stated briefly. High profits that persist are often the result of those private conspiracies against which Smith so eloquently inveighed, or of government dispensations of exclusive privileges.[15] In both cases the result is an impediment to resource mobility upon which the effective functioning of a market economy must be predicated. The alacrity with

[12] Smith (1937, p. 82).

[13] Smith (1937, pp. 249–50).

[14] Smith (1937, p. 250).

[15] Profits must also remain relatively high in some areas to compensate for additional risk or for a disagreeable activity (see Smith 1937, chap. 10, pt. 1). "The keeper of an inn or tavern, who is never master of his own house, and who is exposed to the brutality of every drunkard, exercises neither a very agreeable nor a very creditable business. But there is scarce any common trade in which a small stock yields so great a profit" (p. 101).

which businessmen have entered into such arrangements in the past and their persistence and ingenuity in subverting the disciplining effects of the market are the main reason that the text of *The Wealth of Nations* abounds in phraseology extremely critical of the business community: "the sneaking arts of underling tradesmen"; the "mean and malignant expedients" of merchants and manufacturers; the "clamour and sophistry of merchants and manufacturers"; the "interested sophistry of merchants and manufacturers"; "the mean rapacity, the monopolizing spirit of merchants and manufacturers"; and traders who argue with "all the passionate confidence of interested falsehood."

Smith's criticisms of mercantilism, which take up such a large portion of his book, also issue from the same cause. Smith sees mercantilism as the successful attempt of rapacious businessmen to exploit the machinery of government for their own self-aggrandizement. Such efforts really had their historical origin in the exclusive corporative spirit of privileged groups, which grew up in medieval towns and cities: "Country gentlemen and farmers, dispersed in different parts of the country, cannot so easily combine as merchants and manufacturers, who being collected into towns, and accustomed to that exclusive corporation spirit which prevails in them, naturally endeavour to obtain against all their countrymen, the same exclusive privilege which they generally possess against the inhabitants of their respective towns."[16]

The violence of Smith's polemic against mercantilism lay in the fact that it enabled merchants to better their condition in a manner that did not contribute to the nation's economic welfare. As a result of the dispensation of monopoly grants, of the arbitrary bestowal of "extraordinary privileges" and "extraordinary restraints" upon different sectors of industry by the government, the individual merchant was provided with innumerable opportunities to enrich himself without enriching the nation. Even when legislation is passed with an ostensibly legitimate social purpose in view, the opportunities for profit making are likely to be restructured in such a way as to lead to private enrichment and not social enrichment. Thus, with respect to the herring bounty, Smith sardonically observes: "The bounty to the white herring fishery is a tonnage bounty; and is proportioned to the burden of the ship, not to her diligence or success in the fishery; and it has, I am afraid, been too common for vessels to fit out for the sole purpose of catching, not the fish, but the bounty."[17]

The more interesting part of the answer to my question, however, does not lie in monopolistic barriers or other impediments to the achievement of static efficiency with respect to resource use. Rather, it involves the realm of dynamic change over time and broader influences shaping human behavior. For the growth of trade and commerce—and, in their wake, manufactures—is of course associated historically with the rise of the capitalist class. This class gradually displaces the landlord class, which had previously dominated the European economy and polity and had squandered society's social surplus by maintaining a large army of retainers and by what Smith calls "rustic hospitality." The new goods made available by expanding commerce bring in their wake drastic social and political changes.

---

[16] Smith (1937, p. 429).

[17] Smith (1937, p. 486). Smith adds the following extraordinary bit of accounting: "In the year 1759, when the bounty was at fifty shillings the ton, the whole buss fishery of Scotland brought in only four barrels of sea sticks. In that year each barrel of sea sticks cost government in bounties alone £113 15s; each barrel of merchantable herrings £159 7s. 6d."

> But what all the violence of the feudal institutions could never have effected, the silent and insensible operation of foreign commerce and manufactures gradually brought about. These gradually furnished the great proprietors with something for which they could exchange the whole surplus produce of their lands, and which they could consume themselves without sharing it either with tenants or retainers. All for ourselves, and nothing for other people, seems, in every age of the world, to have been the vile maxim of the masters of mankind. As soon, therefore, as they could find a method of consuming the whole value of their rents themselves, they had no disposition to share them with any other persons. For a pair of diamond buckles perhaps, or for something as frivolous and useless, they exchanged the maintenance, or what is the same thing, the price of the maintenance of a thousand men for a year, and with it the whole weight and authority which it could give them. The buckles, however, were to be all their own, and no other human creature was to have any share of them; whereas in the more ancient method of expence they must have shared with at least a thousand people. With the judges that were to determine the preference, this difference was perfectly decisive; and thus, for the gratification of the most childish, the meanest and the most sordid of all vanities, they gradually bartered their whole power and authority.[18]

The growth of the commercial sector and the increasing control over income flows by the capitalist class are a critical element in Smith's version of economic growth because, whereas the landlord directed society's surplus resources into frivolous, unproductive activities, the capitalist now directs these resources into productive channels. As Smith puts it, "It is the stock that is employed for the sake of profit, which puts into motion the greater part of the useful labour of every society."[19]

---

[18] Smith (1937, pp. 388–89; see also p. 385). For further discussion, see Rosenberg (1968).

Smith's sociological analysis of the rise of capitalism—primarily in book 3 of *The Wealth of Nations*—has been strangely neglected and will, unfortunately, also be neglected here, since it would require a separate paper to treat adequately. A couple of things, however, need to be asserted. By providing a ready market for agricultural products, the growth of commercial and manufacturing towns provides powerful new incentives to the attainment of efficient resource use in agriculture. Furthermore, and for Smith most important, the growth of commerce, by dissolving feudal ties and obligations, makes good government possible for the first time: "Commerce and manufactures gradually introduced order and good government, and with them, the liberty and security of individuals, among the inhabitants of the country, who had before lived almost in a continual state of war with their neighbours, and of servile dependency upon their superiors. This, though it has been the least observed, is by far the most important of their effects."[20]

This good government includes the reduction of crime, which Smith associates with the elimination of the personal ties of dependency of feudalism. He asserts in his *Lectures* that "nothing tends so much to corrupt mankind as dependency, while independency still increases the honesty of the people." And he concludes that "the establishment of commerce and manufactures, which brings about this independency, is the best police for preventing crimes." [21]

Finally, as suggested earlier, the rise of a capitalist class brings an increasing proportion of society's resources— including agriculture itself— under the control of a more efficient class of decision makers.[22] But it is the dynamic aspect of this point that requires emphasis. The growth of commerce *is instrumental in shaping* character, in altering tastes, and in providing new and more powerful incentives. The growth of commerce, by increasing the importance of the capitalist class as compared to large landowners, increases the proportion of those in society devoted to parsimony and frugality ("those who are naturally the most disposed to accumulate"),[23] as compared to those who live lives of indolence and prodigality.[24] Commerce inculcates habits of orderliness, reliability, precision, and painstaking attention to detail. Participation in business enterprise inevitably inculcates certain behavior patterns—in particular, those of "order, economy and attention."[25] Commerce introduces probity and punctuality. But it is important to note that Smith's argument makes these qualities emerge and spread as a direct response to personal self-interest:

---

[19] Smith (1937, p. 249). The structure of feudal society effectively suppressed the possibility of capital accumulation from all classes—albeit in different ways: "Under the feudal constitution there could be very little accumulation of stock, which will appear from considering the situation of those three orders of men, which made up the whole body of the people: the peasants, the landlords, and the merchants. The peasants had leases which depended upon the caprice of their masters; they could never increase in wealth, because the landlord was ready to squeeze it all from them, and therefore they had no motive to acquire it. As little could the landlords increase their wealth, as they lived so indolent a life, and were involved in perpetual wars. The merchants again were oppressed by all ranks, and were not able to secure the product of their industry from rapine and violence. Thus there could be little accumulation of wealth at all; but after the fall of the feudal government these obstacles to industry were removed, and the stock of commodities began gradually to increase" (Smith 1956, p. 220).

[20] Smith (1937, p. 385).

[21] Smith (1956, p. 155).

[22] Smith succinctly lays out the differences in attitude and mentality between the merchant and landowner: "The wealth acquired by the inhabitants of cities was frequently employed in purchasing such lands as were to be sold, of which a great part would frequently be uncultivated. Merchants are commonly ambitious of becoming country gentlemen, and when they do, they are generally the best of all improvers. A merchant is accustomed to employ his money chiefly in profitable projects, whereas a mere country

> Whenever commerce is introduced into any country probity and punctuality always accompany it. These virtues in a rude and barbarous country are almost unknown. Of all the nations in Europe, the Dutch, the most commercial, are the most faithful to their word. The English are more so than the Scotch, but much inferior to the Dutch, and in the remote parts of this country they are far less so than in the commercial parts of it. This is not at all to be imputed to national character, as some pretend; there is no natural reason why an Englishman or a Scotchman should not be as punctual in performing agreements as a Dutchman. It is far more reducible to self-interest, that general principle which regulates the actions of every man, and which leads men to act in a certain manner from views of

---

gentleman is accustomed to employ it chiefly in expence. The one often sees his money go from him and return to him again with a profit; the other, when once he parts with it, very seldom expects to see any more of it. Those different habits naturally affect their temper and disposition in every sort of business. A merchant is commonly a bold; a country gentleman, a timid undertaker. The one is not afraid to lay out at once a large capital upon the improvement of his land, when he has a probable prospect of raising the value of it in proportion to the expence. The other, if he has any capital, which is not always the case, seldom ventures to employ it in this manner. If he improves at all, it is commonly not with a capital, but with what he can save out of his annual revenue" (Smith 1937, pp. 384–85).

[23] Smith (1937, p. 578).

[24] Adam Smith's close friend, David Hume, had said: "(As) the spending of a settled revenue is a way of life entirely without occupation, men have so much need of somewhat to fix and engage them, that pleasures, such as they are, will be the pursuit of the greater part of the landholders, and the prodigals among them will always be more numerous than the misers. In a state, therefore, where there is nothing but a landed interest, as there is little frugality, the borrowers must be very numerous, and the rate of interest must hold proportion to it. The difference depends not on the quantity of money, but on the habits and manners which prevail" (Hume 1955, p. 50).

[25] Smith (1937, p. 385).

> advantage, and is as deeply implanted in an Englishman as a Dutchman. A dealer is afraid of losing his character, and is scrupulous in observing every engagement. When a person makes perhaps twenty contracts in a day, he cannot gain so much by endeavouring to impose on his neighbours, as the very appearance of a cheat would make him lose. When people seldom deal with one another, we find that they are somewhat disposed to cheat, because they can gain more by a smart trick than they can lose by the injury which it does their character. . . . Wherever dealings are frequent, a man does not expect to gain so much by any one contract, as by probity and punctuality in the whole, and a prudent dealer, who is sensible of his real interest, would rather choose to lose what he has a right to, than give any ground for suspicion. Everything of this kind is odious as it is rare. When the greater part of people are merchants, they always bring probity and punctuality into fashion, and these, therefore, are the principal virtues of a commercial nation.[26]

This discussion of the character-forming aspects of a commercial society now provides the basis for our confrontation with the "paradox of high profits" with which this paper is concerned. A commercial society needs to be perceived as a set of institutions, which, although at one level it may be treated as a collection of legally free individuals engaging in free contractual agreements, at another level is an intensely coercive system. By this I mean that, in order to succeed under a system of competitive capitalism, one needs to develop certain characteristics—the characteristics of order, economy, attention, and probity—with which Smith is concerned and which are the qualities essential for success under the unique pressures imposed upon individual participants in the business arena by capitalist institutions. (The capitalist is haunted by the spectre of bankruptcy. "Bankruptcy is perhaps the greatest and most humiliating calamity which can befall an innocent man. The greater part of men, therefore, are sufficiently careful to avoid it. Some, indeed, do not avoid it; as some do not avoid the gallows."[27]) These characteristics, it should be clear, do not come naturally to man. Man does not by nature prefer the active and energetic life to the life of indolence and repose. Indeed, Smith asserts that "it is the interest of every man to live as much at his ease as he can."[28] As a consequence, Smith regards it as axiomatic that "in every profession, the exertion of the greater part of those who exercise it, is always in proportion to the necessity they are under of making that exertion."[29] Landlords "are the only one of the three orders whose revenue costs them neither labour nor care, but comes to them, as it were, of its own accord, and independent of any plan or project of their own."[30] Their characteristic indolence, therefore, is viewed by Smith as "the natural effect of the ease and security of their situation."[31]

---

[26] Smith (1956, pp. 253–55). For Smith's characterization of the "inconveniences" of a commercial society, see Smith (1937, pp. 255–59).

[27] Smith (1937, p. 325).

[28] Smith (1937, p. 718).

But, while the landed classes live a life of indolence, self-indulgence, and ostentation, they are merely doing what other classes would do if they had the opportunity. For, "a man of a large revenue, whatever may be his profession, thinks he ought to live like other men of large revenues; and to spend a great part of his time in festivity, in vanity, and in dissipation."[32] The great virtue of competitive capitalism, from this point of view, is that the intense pressures of the marketplace render such behavior extremely difficult or impossible on the part of the capitalist class. So long as profits are difficult to earn, and so long as competitive pressures keep the rate of profit low, the system itself may be relied upon to force the capitalist to display the traditional virtues of his class. However, high rates of profit, when they persist, constitute evidence that the competitive mechanism is, for whatever reason, not functioning properly. While it is obvious that this has undesirable consequences in terms of resource allocation,[33] it has not been commonly noticed that such easily earned profits had other undesirable consequences, to which Smith attached enormous importance. For

> besides all the bad effects to the country in general, which have already been mentioned as necessarily resulting from a high rate of profit; there is one more fatal, perhaps, than all these

---

[29] Smith (1937, p. 717).

[30] Smith (1937, p. 249).

[31] Smith (1937, p. 249). In speaking of large landed proprietors, Smith remarks: "To improve land with profit, like all other commercial projects, requires an exact attention to small savings and small gains, of which a man born to a great fortune, even though naturally frugal, is very seldom capable. The situation of such a person naturally disposes him to attend rather to ornament which pleases his fancy, than to profit for which he has so little occasion. The elegance of his dress, of his equipage, of his house, and household furniture, are objects which from his infancy he has been accustomed to have some anxiety about. The turn of mind which this habit naturally forms, follows him when he comes to think of the improvement of land. He embellishes perhaps four or five hundred acres in the neighbourhood of his house, at ten times the expence which the land is worth after all his improvements; and finds that if he was to improve his whole estate in the same manner, and he has little taste for any other, he would be bankrupt before he had finished the tenth part of it" (Smith 1937, p. 364).

[32] Smith (1937, p. 766). Similarly, although Smith's statement about landlords who "love to reap where they never sowed" is frequently cited, it is usually cited minus a critical qualification which Smith attaches. "As soon as the land of any country has all become private property, the landlords, *like all other men*, love to reap where they never sowed, and demand a rent even for its natural produce" (Smith 1937, p. 49; emphasis added).

[33] Smith (1937, pp. 564–66).

> put together, but which, if we may judge from experience, is inseparably connected with it. The high rate of profit seems every where to destroy that parsimony which in other circumstances is natural to the character of the merchant. When profits are high, that sober virtue seems to be superfluous, and expensive luxury to suit better the affluence of his situation. But the owners of the great mercantile capitals are necessarily the leaders and conductors of the whole industry of every nation, and their example has a much greater influence upon the manners of the whole industrious part of it than that of any other order of men. If his employer is attentive and parsimonious, the workman is very likely to be so too; but if the master is dissolute and disorderly, the servant who shapes his work according to the pattern which his master prescribes to him, will shape his life too according to the example which he sets him. Accumulation is thus prevented in the hands of all those who are naturally the most disposed to accumulate; and the funds destined for the maintenance of productive labour receive no augmentation from the revenue of those who ought naturally to augment them the most. The capital of the country, instead of increasing, gradually dwindles away, and the quantity of productive labour maintained in it grows every day less and less. Have the exorbitant profits of the merchants of Cadiz and Lisbon augmented the capital of Spain and Portugal? Have they alleviated the poverty, have they promoted the industry of those two beggarly countries? Such has been the tone of mercantile expence in those two trading cities, that those exorbitant profits, far from augmenting the general capital of the country, seem scarce to have been sufficient to keep up the capitals upon which they were made. . . . Compare the mercantile manners of Cadiz and Lisbon with those of Amsterdam, and you will be sensible how differently the conduct and character of merchants are affected by the high and by the low profits of stock. . . . Light come light go, says the proverb; and the ordinary tone of expence seems every where to be regulated, not so much according to the real ability of spending as to the supposed facility of getting money to spend.[34]

It is only the force of competition, apparently, which can be relied upon to keep the capitalist from behaving like an extravagant landowner. This is so because a major determinant of economic behavior is the ease or difficulty involved in the earning of income. While it may be going too far to suggest that, although Smith did not subscribe to a backward-sloping supply curve for labor, he *did* subscribe to it for the capitalist, he does believe that a rise in the rate of profit will reduce the quality, if not the supply, of capitalist effort.

---

[34] Smith (1937, pp. 578-79).

It is true that the barbs which Smith directed at the wealthy usually have large landowners as their target. But there is a good historical reason for this. When Smith wrote, in the middle of the eighteenth century, the landowning classes still thoroughly dominated English society and provided far more conspicuous targets for his attack on great wealth than did the rising class of merchants and manufacturers. But it should be abundantly clear from what has preceded that Smith's sharp invective against the "indolence and vanity of the rich"[35] is not, in principle or intention, confined to any single class in society. Rather, these are characteristics which are attached to the possessors of wealth, from whatever source that wealth is derived, because such possession conditions its owners in highly predictable ways.[36]

This remark brings me to what is both my final point and perhaps a new paradox to replace the one which I have attempted to resolve. I have argued that Smith's hostility to high profits is rooted in his belief that such profits dull the edges of capitalist performance—as in Cadiz and Lisbon—both by dulling his incentive and capacities as an earner of income and by destroying his frugality in disposing of that income. The trouble—and the paradox—of high profits is that the attainment of wealth corrupts the forces leading to the generation of wealth—as is obviously the case with the large landowner. Therefore a recurring theme of the book bearing the title *An Inquiry into the Nature and Causes of the Wealth of Nations* is that, at least on the individual level, the easy attainment of great wealth is likely to destroy the individual's capacity to contribute to the wealth of nations.[37] In this respect, the supreme and essential virtue of competition is that, while it permits the attainment of modest wealth, it places the easy amassing of great wealth virtually beyond reach. It may fairly be said, therefore, that although Adam Smith certainly does not celebrate the social role of the individual capitalist, he does indeed celebrate the role of the capitalist system—or, more precisely, the role of competitive capitalism.

Yet, with all of Smith's preoccupation with the wealth of nations, he also believes that the pursuit of wealth does not take place for the direct gratification or utilitarian purposes provided by an abundance of worldly goods, but rather because the possession of such goods brings their owner the high esteem and approbation of his fellow man.[38] That paradox—and surely the insistence upon the relative unimportance of the wealth of nations by the author of *The Wealth of Nations* deserves to be called a paradox—has to be pursued through Smith's earlier work, *The Theory of Moral Sentiments*. But that is another story.

---

[35] Smith (1937, p. 683).

[36] See Rosenberg (1960).

[37] Smith also noted the inverse correlation between income level and human fertility: "A half-starved Highland woman frequently bears more than twenty children, while a pampered fine lady is often incapable of bearing any, and is generally exhausted by two or three. Barrenness, so frequent among women of fashion, is very rare among those of inferior station. Luxury in the fair sex, while it inflames perhaps the passion for enjoyment, seems always to weaken, and frequently to destroy altogether, the powers of generation" (Smith 1937, p. 79). It is curious that Malthus never examined the important implications of this statement for his theory of population.

References

Coats, A. W. "Changing Attitudes to Labour in the Mid-eighteenth Century." *Econ. Hist. Rev. 11* (August 1958): 35–51.

Furniss, Edgar S. *The Position of the Laborer in a System of Nationalism.* New York: Kelley & Millman, 1957.

Hume, David. *Writings on Economics.* Edited by Eugene Rotwein. Madison: Univ. Wisconsin Press, 1955.

Meek, Ronald. "Adam Smith and the Classical Concept of Profit." *Scottish J. Polit. Econ. 1* (June 1954): 138–53.

Rosenberg, Nathan. "Some Institutional Aspect of the *Wealth of Nations.*" *J.P.E.* 68 (December 1960): 557–70.

———. "Adam Smith, Consumer Tastes, and Economic Growth." *J.P.E.* 76 (May/June 1968): 361–74.

Smith, Adam. *Wealth of Nations.* New York: Modern Library, 1937.

———. *Lectures on Justice, Police, Revenue, and Arms.* New York: Kelley & Millman, 1956.

———. "The Theory of Moral Sentiments." In *Essays, Philosophical and Literary.* London: 1880.

Temple, Sir William. *Vindication of Commerce and the Arts.* London: 1786.

Young, Arthur. *Farmer's Tour through the East of England.* vol. 4. London: 1771.

[38] "(W)hat are the advantages which we propose by the great purpose of human life which we call bettering our condition? To be observed, to be attended to, to be taken notice of with sympathy, complacency, and approbation, are all the advantages which we can propose to derive from it. It is the vanity, not the ease or the pleasure, which interests us. But vanity is always founded upon the belief of our being the object of attention and approbation. The rich man glories in his riches, because he feels that they naturally draw upon him the attention of the world, and that mankind are disposed to go along with him in all those agreeable emotions with which the advantages of his situation so readily inspire him. At the thought of this, his heart seems to swell and dilate itself within him, and he is fonder of his wealth upon this account, than for all the other advantages it procures him" (Smith 1880, p. 48). Smith makes this point almost aphoristically in the *Wealth of Nations* when he asserts that "with the greater part of rich people, the chief enjoyment of riches consists in the parade of riches" (1937, p. 172); see also Rosenberg (1968, pp. 364–67).

Tariffs and the Distribution of Income: The Importance of Factor Specificity, Substitutability, and Intensity in the Short and Long Run

---

Michael Mussa

University of Rochester

This paper analyzes the effects of changes in relative commodity prices on the distribution of income among factors of production in the context of two models of a simple, two-good economy. In the first model capital is treated as a specific factor in each industry, with labor mobile between industries. The assumption of specificity determines the direction of factor income changes, with magnitudes depending on substitutability between factors and on intensities of factor use within the two industries. In the second model, capital is viewed as a quasi-fixed factor. For the short run, this model is identical to the model first considered. For the long run, this model is identical to the Stolper-Samuelson model in which the direction and magnitude of factor income changes depend solely on relative factor intensities. The difference between the short-run and long-run determinants of changes in factor incomes gives rise to a conflict between factor owners' short-run and long-run interests.

Introduction

A primary effect and principal objective of commercial policy is frequently to protect or enhance the incomes of specially favored groups. The modern theory of the effects of commercial policy on the distribution of income, the Stolper-Samuelson theory, however, is not adequate for analyzing much of the clamor for protection. This theory envisions an economy in which factors of production are costlessly and instantaneously mobile between productive activities and emphasizes the effects of differences in relative factor intensities. The theory suggests that if automobile

I would like to acknowledge the many useful comments and suggestions which were made by my colleagues Rudiger Dornbusch, Ronald Jones, and Sherwin Rosen. I would also like to acknowledge the financial support for this research which was provided by the National Science Foundation through the MSSB Summer Workshop on the Application of General Equilibrium Models to Economic History.

[*Journal of Political Economy*, 1974, vol. 82, no. 6]

© 1974 by The University of Chicago. All rights reserved.

manufacturing is capital intensive, then General Motors will favor increases in the tariff on automobiles and the United Auto Workers will oppose such increases; and, if textile manufacturing is labor intensive, then textile workers will favor quotas on textile imports, and domestic textile manufacturers will oppose them. What the theory neglects is that, in the short run, at least, factors tend to be specific to particular uses.[1] The stamping machines and assembly lines of automobile manufacturers are not costlessly transformable into the looms and weaving mills of textile manufacturers. The skills of a tool-and-die maker in Detroit are not instantaneously interchangeable with those of a master weaver in North Carolina.

The objective of this paper, however, is not to criticize the Stolper-Samuelson theory, but rather to integrate it with an earlier "Marshallian" approach. An integration of the two approaches is desirable because each lays emphasis on a different set of important phenomena. The "Marshallian" approach emphasizes the distinction between the short run and the long run.[2] For the short run, it focuses on specific factors and on the degree of substitutability between these factors and highly mobile factors. For the long run, it recognizes that factors that are specific in the short run often can be moved to alternative uses. The analysis of the long run is complemented by the Stolper-Samuelson theory, which assumes complete mobility of factors and focuses on general equilibrium interactions between factor endowments and the factor intensities of different productive processes.

The two approaches will be integrated within the context of two related models of a simple, two-good economy.[3] In the first model, capital will be treated as a fixed factor that is specific to the industry in which it is used, while labor will be assumed to be free to move between industries. These assumptions about factor mobility will suffice to determine the direction of changes in factor incomes in response to changes in output prices: the income of capital in each industry rises more than proportionately with increases in the price of its own output and falls with increases in the price of the other output; the income of labor rises less than proportionately with increases in either output price. The magnitude of factor income changes, however, will be shown to depend on the substitutability between labor and capital in each industry and on the factor intensities of the two industries.

The second model will preserve the structure of the first, except that capital will be treated as a quasi-fixed factor that is specific to a given industry at a moment of time, but free to move between industries in the long run. For the short run, this model is identical to the first model. For the long run, it is identical to the Stolper-Samuelson model. By using known properties of the Stolper-Samuelson model, it will be shown that the short-run and long-run determinants of the behavior of factor incomes are very different and that these differences necessarily imply a conflict between factor owners' short-run and long-run interests.

---

[1] In the introduction to their classic article, Stolper and Samuelson (1941) take note of the potential importance of specific factors, but do not integrate such factors into their formal analysis. In the more recent literature, specific factors have received relatively little attention. A notable exception is Jones (1971).

[2] For an excellent summary of the "Marshallian" approach to the analysis of the effects of commodity price changes on factor incomes, see Pigou (1906, especially pp. 55–59).

[3] In a recent paper, brought to my attention by the editor of this *Journal*, Mayer (1974) analyzes similar models, focusing on a slightly different set of issues.

I. Factor Specificity and the Direction of Income Changes

Consider a two-commodity, three-factor model with the following properties. (1) There is a single mobile factor, labor, which is used in the production of both commodities, X and Z. (2) There are two specific factors, capital in X and capital in Z, which are used only in their respective industries, and which are in fixed supply to those industries. (3) The production functions for the two commodities are each linear homogeneous in their respective inputs and have the standard neoclassical properties of differentiability and of positive and declining marginal physical products for each of the inputs, specifically,

$$X = F(L_X, K_X), \tag{1}$$

$$Z = G(L_Z, K_Z). \tag{2}$$

(4) The total quantity of labor used in both industries is equal to the fixed aggregate supply of labor; that is,

$$L_X + L_Z = \bar{L}. \tag{3}$$

Given an initial relative price of X in terms of Z (within the range of nonspecialization), say P_X^0, the distribution of the labor force, the level of the wage rate, and the income of capital in X and capital in Z may be determined with the aid of figure 1. The length of the horizontal axis is equal to the total supply of labor, $\bar{L}$. The vertical axes passing through 0_X and 0_Z measure the wage of labor in terms of units of Z per unit of labor. The curve labeled $VMPL_X(P_X^0)$ is plotted relative to the origin 0_X and shows the value of the marginal product of labor in X as a function of L_X. The vertical position of this curve depends on the given value of the relative price of X, P_X^0. The curve labeled $VMPL_Z$ is plotted relative to the origin 0_Z and shows the value of the marginal product of labor in Z as a function of L_Z. Its position is independent of the commodity-price ratio. The point of intersection, (L^0, w^0), determines the equilibrium distribution of the labor force and the equilibrium value of the wage rate. The income of labor in X and the income of labor in Z are shown by the corresponding rectangular areas under the wage line, $w = w^0$. The incomes of the two types of capital are shown by the corresponding

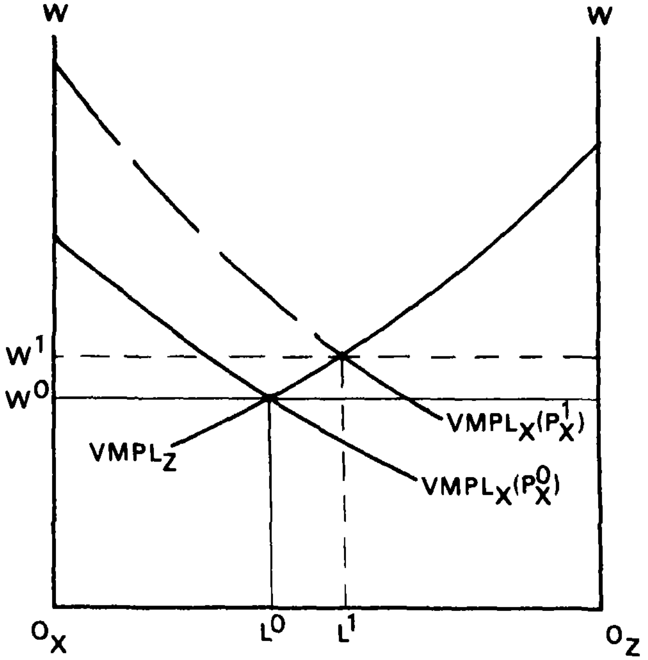


FIG. 1.—Equilibrium wage rate and allocation of the labor force

triangular regions between the wage line and the respective value of marginal product curves.

An increase in P_X, say from P_X^0 to P_X^1, shifts the $VMPL_X$ curve proportionately upward, resulting in a new equilibrium at (L^1, w^1). The increase in the wage rate (measured in terms of Z), however, is less than the proportionate increase in the relative price of X in terms of Z. Labor, the mobile factor, gains in terms of Z but loses in terms of X. In contrast, capital in X gains in terms of both commodities, and capital in Z loses in terms of both commodities.[4] The loss to capital in Z is apparent from the fact that the increase in the wage rate reduces the size of the triangular region which corresponds to the income to capital in Z. The gain to capital in X can be seen by considering only the units of X which were (and still are) produced by the original amount of labor, L^0. The value of these units of X rises proportionately with the increase in P_X. The cost of the labor employed in producing these units rises, but by less, proportionately, than the increase in P_X. Hence, the return to capital, on just these original units, must rise by more, proportionately, than the increase in P_X. Since capital in X also earns a positive amount on the additional units which are produced, it follows that capital in X must gain in terms of both goods.

---

[4] Marshall (1961, p. 664), summarizes these results in the following words: "The employer stands as a buffer between the buyer of goods and all the various classes of labor by which they are made. He receives the whole price of the one and pays the whole price of the others. Fluctuations of his profits go with fluctuations of the prices of the things he sells, and are more extensive: while those of wages of his employees are less extensive."

The assumption of specific capital and mobile labor determines the direction of factor income changes in response to commodity-price changes. The property of "specificity," however, should not be conceived of as a wholly technological matter. If P_X is so high that all labor is specialized in the production of X, then the income of labor will respond in exactly the same way as the income of capital specific to X to an increase in P_X: both will increase proportionately with the increase in P_X. "Specificity" is an economic, as well as a technological, matter.

The results of the present model concerning the responses of factor incomes to changes in commodity prices can be interpreted in terms of the "Marshallian" concept of "rents". "Rents" are payments to factors of production in excess of what they can earn in their best alternative use. Since capital in each industry is specific to that industry, the incomes of both types of capital can be thought of as rents. In a sense, the payments to labor can also be thought of as "rents." For while the supply of labor to each industry is not absolutely fixed, the total supply of labor to both industries is fixed, and the supply of labor to any one industry is less than infinitely elastic. In fact, the supply curve of labor facing the X industry is precisely the $VMPL_Z$ curve looked at from the origin 0_X. It is because this supply curve is positively sloped that an increase in the demand for labor in X cannot be absorbed without an increase in the wage rate.

II. Substitutability, Intensity, and the Magnitude of Income Changes

In the model of the preceding section, the direction of factor income changes is completely determined by the assumption of specific capital and mobile labor. The magnitudes of factor income changes, however, depend on the substitutability between labor and capital in the two industries and on the intensities with which the two factors are used in the two industries. In this section, we will examine the magnitude of the factor income change, first for labor, then for the two specific types of capital, and finally for capital as a whole.

A. The Income of Labor

A formal expression for the change in the equilibrium wage rate may be obtained by using the labor-market equilibrium condition[5]

$$L_X^d(w/P_X, K_X) + L_Z^d(w, K_Z) = \bar{L}, \tag{4}$$

---

[5] An alternative approach to deriving a number of the results discussed in this section is given in Jones (1971).

where L_X^d and L_Z^d are the labor-demand functions for the two industries (the inverse of the *VMPL* functions). Differentiating equation (4), holding K_X and K_Z constant, we obtain

$$\hat{w} = \frac{\lambda_{LX}\zeta_X}{\lambda_{LX}\zeta_X + \lambda_{LZ}\zeta_Z}\hat{P}_X \equiv \eta\hat{P}_X, \tag{5}$$

where $\hat{w} \equiv dw/w$, $\hat{P}_X \equiv dP_X/P_X$,[6] and λ_{Li} is the fraction of the labor force employed in industry i, and ζ_i is the elasticity of demand for labor in industry i. The magnitude η, which is defined in (5), is the elasticity of the wage rate with respect to the relative price of X. By making use of the fact that the elasticity of demand for labor in each industry can be written as

$$\zeta_i = \frac{\sigma_i}{1 - \theta_{Li}}, \tag{6}$$

where σ_i is the elasticity of substitution between labor and capital in industry i, and θ_{Li} is the distributive share of labor in the value of output in industry i, the elasticity η may be rewritten as[7]

$$\eta = \frac{\lambda_{LX}\left(\frac{\sigma_X}{1 - \theta_{LX}}\right)}{\lambda_{LX}\left(\frac{\sigma_X}{1 - \theta_{LX}}\right) + \lambda_{LZ}\left(\frac{\sigma_Z}{1 - \theta_{LZ}}\right)}. \tag{7}$$

The result (6) shows the importance of factor substitution and factor intensity (as measured by distributional shares) for the responsiveness of the wage rate to changes in P_X. Holding the λ's and θ's constant, when σ_X is large and σ_Z is small, the behavior of the wage rate follows very closely the behavior of P_X. Geometrically, the reason for this is that the $VMPL_X$ curve is very flat, while the $VMPL_Z$ curve is very steep; vertical shifts in the $VMPL_X$ curve, therefore, result in virtually equal changes in the equilibrium wage rate. The same result obtains when θ_{LX} is large and θ_{LZ} is small, and for the same reason. Factor substitutability and factor intensity affect the responsiveness of the wage rate in the same way because both have similar effects on the elasticities of demand for labor in the two industries. From equation (5) it is apparent that these elasticities are the determinants of the elasticity of the wage rate with respect to P_X.

---

[6] We will use a caret to indicate the operation of taking a percentage change.

[7] I am indebted to Carlos Rodriguez for suggesting the rewriting of eq. (5) in the form of eq. (7).

B. The Income of Specific Capital

The income of capital in each industry is determined as a residual after payments to labor:

$$Y_{KX} = P_x X - wL_X; \qquad Y_{KZ} = Z - wL_Z. \tag{8}$$

By differentiating these equations, making use of the fact that the wage rate is equal to the value of the marginal product of labor in each industry, and expressing the results in elasticity form, it follows that

$$\hat{Y}_{KX} = \left(\frac{1}{\theta_{KX}}\right)(1 - \theta_{LX}\eta)\hat{P}_X, \tag{9}$$

$$\hat{Y}_{KZ} = \left(\frac{-1}{\theta_{KZ}}\right)(\theta_{LZ}\eta)\hat{P}_X, \tag{10}$$

where $\theta_{KX} \equiv Y_{KX}/P_x X = 1 - \theta_{LX}$ and $\theta_{KZ} \equiv Y_{KZ}/Z = 1 - \theta_{LZ}$. The slight asymmetry between (9) and (10) arises because Z has been taken as the numeraire.

The results in equations (9) and (10) demonstrate the importance of factor intensities and, indirectly, of factor substitutability for the magnitudes of $\hat{Y}_{KX}$ and $\hat{Y}_{KZ}$. The degree of substitutability between labor and capital in the two industries affects these results through its influence on η. From (7) it follows that an increase in σ_X or a decrease in σ_Z will decrease the absolute value $\hat{Y}_{KX}$ and increase the absolute value of $\hat{Y}_{KZ}$. The essential symmetry of this result may be seen more clearly when it is restated as follows: the greater the elasticity of substitution between labor and capital in a given industry, relative to what it is in the other industry, the more closely the income of labor and the income of capital in that industry mirror the behavior of the price of that industry's output. When σ_X is relatively large, both $\hat{w}/\hat{P}_X$ and $\hat{Y}_{KX}/\hat{P}_X$ tend to be close to unity.

The value of η can be thought of as determining the degree of "effective protection" which is afforded to the two types of specific capital. From the viewpoint of the owners of capital, labor is purchased input, just as material inputs would be purchased inputs if intermediate goods were used in production. If the price of the purchased input (i.e., the wage rate) were to remain constant when P_X rises, then the degree of "effective protection" afforded to capital in X (as measured by $\hat{Y}_{KX}$) would equal $\hat{P}_X/\theta_{KX}$. The fact that the wage rate rises when P_x rises means that the degree of "effective protection" which is afforded to capital in X is smaller than $\hat{P}_X/\theta_{KX}$ by the amount $(\theta_{LX}/\theta_{KX})\hat{w}$. From the standpoint of the owners of capital in Z, the increase in the wage rate which is induced by an increase in P_X means that capital in Z suffers from "negative effective protection" to the extent of $(\theta_{LZ}/\theta_{KZ})\hat{w}$.

The importance of factor intensities is also apparent in equations (9) and (10). The inverse of capital's share in each industry determines the "leverage" of that type of capital with respect to changes in relative commodity prices. When capital's share is unity, the income of capital moves in proportion with the price of its output. As capital's share decreases, the percentage change in the income of capital increases with the inverse of capital's share.

Formally, differentiating (9) and (10) with respect to θ_{KX} and θ_{KZ}, holding η constant, we obtain

$$\frac{d\hat{Y}_{KX}}{d\theta_{KX}} = \frac{-(1-\eta)}{(\theta_{KX})^2}\hat{P}_X, \tag{11}$$

$$\frac{d\hat{Y}_{KZ}}{d\theta_{KZ}} = \frac{\eta}{(\theta_{KZ})^2}\hat{P}_X. \tag{12}$$

It follows that the greater the intensity of capital in an industry, the more closely the income of that capital mirrors the price of its output (e.g., the greater θ_{KX}, the closer is $\hat{Y}_{KX}/\hat{P}_X$ to unity). On the other hand, the leverage effect says that the smaller the intensity of capital in a particular industry, the more strongly the income of that capital responds the price of its output (e.g., the smaller θ_{KX}, the larger is $\hat{Y}_{KX}/\hat{P}_X$). However, both of these conclusions may be reversed if the indirect effects of changes in factor intensities on the value of η (holding σ_X and σ_Z constant) are incorporated into the results (11) and (12).

C. The Income of Capital as a Whole

The question of what happens to the income of capital as a whole when P_x rises is likely to be interesting if ownership of the two types of capital is widely diversified within the capital-holding class. Diversification is to be expected if the claims to the ownership of the two types of capital are fairly easily marketable. In such a circumstance, the negative correlation between the incomes of the two types of capital which will be produced by relative price changes will provide a strong incentive to portfolio diversification and, hence, lead the capital-holding class to be more concerned with the income of capital as a whole rather than with the income of either specific type of capital.

The income of capital as a whole, Y_K, is given by

$$Y_K = Y - wL = (P_X X + Z) - wL. \tag{13}$$

If we differentiate this equation, making use of the fact that $dY = XdP_X$, and express the result in elasticity form, it follows that

$$\begin{aligned} \hat{Y}_k &= \frac{1}{\theta_K}(\hat{Y} - \theta_L \hat{w}) \\ &= \frac{1}{\theta_K}(\beta_X - \theta_L \eta)\hat{P}_X, \end{aligned} \tag{14}$$

where $\theta_K \equiv Y_K/Y$, $\theta_L \equiv Y_L/Y$, and $\beta_X \equiv P_X X/Y$. From equation (14) it is apparent that the income of capital as a whole may either rise or fall as a result of an increase in P_X. A decline is likely when the aggregate share of labor, θ_L, and the elasticity of the wage rate, η, are large, and when the share of X in the value of total output is small.

The second line of equation (14) is identical to the result given in equation (9) for the value of $\hat{Y}_{KX}$, except that the fraction β_X has replaced the number 1. As in the case of Y_{KX}, the inverse of θ_K determines the leverage of capital as a whole with respect to changes in P_X, and the value of η determines the degree of effective protection which is afforded to capital as a whole. The importance of the elasticities of substitution and distributional shares within industries for the value of $\hat{Y}_K$ is apparent from the roles which these parameters play in the determination of η.

III. Capital as a Quasi-fixed Factor

The analysis of the last two sections can be given an added dimension by viewing capital as a quasi-fixed factor.[8] In the short run, each unit of capital is specific to the industry in which it is engaged. The short-run effects of a change in relative commodity prices are exactly as described in the last two sections. Over time, however, capital can move from one industry to another in search of its highest reward. Long-run equilibrium requires that the distribution of the capital stock be such that the value of the marginal product of capital is equal in the two industries.

A. The Effects of a Shift of Capital

Starting from an initial long-run equilibrium, an increase in P_X leads to a short-run equilibrium in which the value of the marginal product of capital in X exceeds the value of the marginal product of capital in Z. Capital moves from Z to X. To determine the effect on the wage rate of a redistribution of the capital stock, differentiate the short-run equilibrium condition (4) with respect to K_X and K_Z, impose $dK_Z = -dK_X$, and make use of the fact that linear homogeneity of the production functions implies $\partial L_i^d/\partial K_i = L_i/K_i$, to conclude that

---

[8] For many applied problems, it is appropriate to think of labor, as well as capital, as a quasi-fixed factor. See Oi (1962) on the subject of labor as a quasi-fixed factor.

$$\hat{w} = \frac{-1}{\lambda_X \zeta_X + \lambda_Z \zeta_Z} [(L_X/K_X) - (L_Z/K_Z)](dK_X/\bar{L}). \tag{15}$$

Further, if we make use of (6), this result can be rewritten as

$$\hat{w} = \frac{-(\theta_{LX} - \theta_{LZ})}{\lambda_X \sigma_X (1 - \theta_{LZ}) + \lambda_Z \sigma_Z (1 - \theta_{LX})} (r/w)(dK_X/\bar{L}), \tag{16}$$

where r is the rental rate on capital at the initial long-run equilibrium (measured in terms of Z).[9]

Since the elasticities which appear in the denominators of equations (15) and (16) are negative, it follows that the direction of the change in the wage rate which results from the shift of an initial unit of capital from Z to X depends only on the long-run relative factor intensities of the two industries, as measured by either the factor ratios or by distributional factor shares. As suggested by the Stolper-Samuelson theorem, the wage rate rises if and only if X production is relatively labor intensive.[10] Note, however, that the magnitude of the change in the wage rate which results from the shift of an initial unit of capital also depends on the degree of substitutability between labor and capital in the two industries.

Geometrically, what is happening is that the shift of capital shifts both the $VMPL_X$ curve and the $VMPL_Z$ curve of figure 1 to the right. The wage rate rises if and only if the horizontal shift of the $VMPL_X$ curve exceeds the horizontal shift of the $VMPL_Z$ curve, at $w = w^1$. The latter condition is satisfied if and only if the L_X^1/K_X^0 is greater than $L_Z^1 K_Z^0$.[11]

B. Long-run Equilibrium and the Stolper-Samuelson Model

In the long run, capital is completely mobile between industries, and the model presented in Section I reduces to the Stolper-Samuelson model in which direction of change in factor incomes induced by an increase in P_X depends only on the (long-run) relative factor intensities of X and Z.

---

[9] Recall that we are considering small changes, starting at a position of initial long-run equilibrium.

[10] If X production is capital intensive in the long run, then a large increase in P_x may result in a short-run reversal of the relative factor proportions (which is not the same thing as a "factor intensity reversal" in the Stolper-Samuelson model). In this event, the wage rate will initially rise as capital begins to shift. However, once the original relationship between factor proportions has been restored, the wage rate will start to fall and must eventually fall below its previous long-run equilibrium value.

[11] From linear homogeneity of the two production functions, it follows that the horizontal shift of the $VMPL$ curve which results from the shift of one unit of capital is proportional to the labor-capital ratios in the respective industries.

The factor that is used relatively intensively in the production of X gains in terms of both goods, and the other factor loses in terms of both goods. Further, from Jones's results it is also known that, at least for small increases in P_X, the magnitude of factor income changes in the Stolper-Samuelson model depends only on factor intensities.[12]

Formally, differentiating the zero-profit restrictions for the two industries (which requires that the weighted sums of factor prices equal the respective output prices, where the weights are the amounts of labor and capital used to produce a unit of output) and expressing the results in terms of percentage changes, it follows that

$$\theta_{LX}\hat{w} + \theta_{KX}\hat{r} = \hat{P}_X, \tag{17}$$

$$\theta_{LZ}\hat{w} + \theta_{KZ}\hat{r} = 0. \tag{18}$$

Solving for $\hat{w}$ and $\hat{r}$ yields

$$\hat{w} = \frac{\theta_{KZ}}{\theta_{LX} - \theta_{LZ}} \hat{P}_X, \tag{19}$$

$$\hat{r} = \frac{-\theta_{LZ}}{\theta_{LX} - \theta_{LZ}} \hat{P}_X. \tag{20}$$

These results reveal great differences among the parameters that are important for determining the direction and magnitude of factor income changes in the long run, after redistribution of the capital stock is complete; and those that are relevant in the short run, in which capital is assumed to be fixed; or during the period of adjustment of the capital stock. In the short run, the direction of factor income changes is wholly determined by the assumption that labor is mobile while capital is immobile: labor gains in terms of Z and loses in terms of X when P_X rises, while capital in X gains in terms of both goods and capital in Z loses in terms of both goods. The magnitude of short-run changes in factor incomes depends on the degree of substitutability between capital and labor in the two industries and on the factor intensities in the two industries. During the period of adjustment, directions of income changes depend only on factor intensities, but magnitudes also depend on substitutability. In the long run, the direction of factor income changes depends only on the difference in factor intensities between the two industries, $\theta_{LX} - \theta_{LZ}$, and the magnitudes of factor income changes are independent of the degree of substitutability between capital and labor in either industry.[13] Further, the Stolper-Samuelson model, which is relevant for long-run analysis, has a "knife edge" property which has no counterpart in the short run. From (19) and (20) it follows that for a given value of $\hat{P}_X$ the magnitude of the implied factor imcome changes becomes very large as the factor intensities of the two industries come close together. Thus, capital gains from an increase in P_X provided that X is relatively capital intensive, but capital gains the most when the intensity differential is the smallest.

---

[12] See Jones (1965) for a more complete explanation of the derivation of the results discussed in this paragraph.

[13] This is apparent from the fact that the elasticities of substitution do not enter into (19) and (20). These results are unaffected by assuming a fixed-coefficients technology for which the elasticity of substitution in each industry is zero.

Conclusion: The Conflict between Short-Run and Long-Run Interests

These differences between the determinants of short-run and long-run changes in factor incomes imply a conflict between factor owners' short-run and long-run interests with respect to commodity prices. Consider any policy which results in an increase in P_X. If X production is relatively capital intensive, then laborers gain in the short run in terms of X but lose in the long run in terms of both goods, and the owners of capital initially employed in Z lose in the short run in terms of both goods but gain in the long run in terms of both goods. On the other hand, if X production is relatively labor intensive, then, even though capital initially employed in X gains in terms of both goods in the short run, it loses in terms of both goods in the long run; and, even though labor loses in terms of X in the short-run, it gains in terms of both goods in the long run. Hence, no matter what the factor intensities of the two industries, there must be at least one factor whose long-run interest runs counter to its short-run interest. Further, if the factor intensities of the two industries are similar, then, in the long run, large changes in factor incomes will be required to accommodate small changes in commodity prices. Not only will some factor's short-run interest differ from its long-run interest, but it will differ a great deal.

Appendix

Generalization of the Results of Section II to an N Commodity, $N + 1$ Factor Model

The results which were obtained for the simple, two-good, three-factor model in Section II generalize in a straightforward way to a model with N commodities, each produced with its own specific capital and mobile labor.[14] Equilibrium in the distribution of the labor force requires that the value of the marginal product of labor be equated across industries. If a government policy produces a vector of commodity price changes $(dP_1, dP_2, \ldots, dP_N)$, then, differentiating the equilibrium condition and expressing the result in the form of percentage changes, it follows that

---

[14] The results presented in this Appendix can also be applied to the case where there are many different firms, each with its specific capital, all producing two or more commodities.

$$\hat{w} = \sum_{i=1}^{N} \mu_i P_i, \tag{21}$$

where $\hat{P}_i \equiv dP_i/P_i$ and

$$\mu_i = \frac{\lambda_{Li}\zeta_{Li}}{\sum_{j=1}^{n} \lambda_{Lj}\zeta_{Lj}} = \frac{[\lambda_{Li}/(1 - \theta_{Li})]\sigma_i}{\sum_{j=1}^{n} [\lambda_{Li}/(1 - \theta_{Lj})]\sigma_j}, \tag{22}$$

where $\lambda_{Li} \equiv L_i/\bar{L}$, $\theta_{Li} \equiv wL_i/P_iX_i$, ζ_{Li} is the elasticity of demand for labor in industry i, σ_i is the elasticity of substitution between labor and capital in industry i, and X_i is the output of good i. Further, using the same procedures as in Section II, it follows that

$$\hat{Y}_{Ki} = \left(\frac{1}{\theta_{Ki}}\right)(\hat{P}_i - \theta_{Li}\hat{w}), \tag{23}$$

$$\hat{Y}_K = \left(\frac{1}{\theta_K}\right)(\hat{Y} - \theta_L\hat{w}), \tag{24}$$

where $\theta_{Ki} \equiv Y_{Ki}/P_iX_i$, $\theta_K \equiv Y_K/Y$, $\theta_L \equiv Y_L/Y = w\bar{L}/Y$, $Y \equiv \sum_{i=1}^{N} P_iX_i$, $\beta_i \equiv P_iX_i/Y$, and

$$\hat{Y} = \sum_{i=1}^{N} \beta_i\hat{P}_i \tag{25}$$

Inspection reveals that all of these results are direct analogues of the results for the two-good case and reduce to the two-good results when $N = 2$.

References

Jones, R. W. "The Structure of Simple General Equilibrium Models." *J.P.E.* 73, no. 6 (December 1965): 557–72.

———. "A Three-Factor Model in Theory, Trade, and History." In *Trade, the Balance of Payments, and Growth*, edited by J. N. Bhagwati. Amsterdam: North-Holland, 1971.

Marshall, A. *Principles*. 9th ed. London: Macmillan, 1961.

Mayer, W. "Short-Run and Long-Run Equilibrium for a Small Open Economy." *J.P.E.* 82, no. 5 (September/October 1974): 955–68.

Oi, W. "Labor as a Quasi-fixed Factor." *J.P.E.* 70, no. 6 (December 1962): 538–55.

Pigou, A. *Protective and Preferential Import Duties*. London: Macmillan, 1906. Reprinted in London School Econ. and Polit. Sci. Series Reprints of Scarce Works on Polit. Econ., no. 2. London: London School Econ.

Stolper, W., and Samuelson, P. "Protection and Real Wages." *Rev. Econ. Studies* 9, no. 1 (November 1941): 58–73.

The Stability of the Demand-for-Money Function in the United States 1901–1965

Mohsin S. Khan

International Monetary Fund

The debate between the Keynesians and monetarists regarding the role of money in determining economic activity and the relative importance of monetary versus fiscal policy can be considered in the framework of the relative stability of the demand-for-money and expenditure functions. This paper tests for the structural stability of the demand-for-money function in the United States over the period 1901–65 utilizing a technique developed by R. L. Brown and J. Durbin. It is shown that the demand-for-money function does not appear to have shifted significantly during the period of study.

I. Introduction

In his well-known survey of monetary theory, Johnson (1962) listed three outstanding issues in the theory of the demand for money: (1) the relevant definition of the money aggregate; (2) the choice of variables explaining movements in this aggregate; and (3) the stability of the functional relationship between money and the explanatory variables. Most empirical studies on the demand for money have concentrated on the first two of these issues, and there are very few studies explicitly concerned with analyzing the question of whether the relationship is stable; the term "stability" is defined in this context in the statistical sense of the estimated coefficients of the explanatory variables remaining constant over time.[1]

I am indebted to Professor H. G. Johnson, Professor R. E. Quandt, P. B. Kenny, and two anonymous referees. The views expressed in this paper are mine and do not necessarily reflect the views of the International Monetary Fund.

[1] The exceptions to this are the studies by Meltzer (1963), Courchene and Shapiro (1964), and Laidler (1971). This interpretation of stability has been used by Friedman (1956). In the Friedman framework, a demand-for-money function should be formulated such that it remains stable with respect to a given number of predetermined variables under differing institutional arrangements, changes in social and political environment, etc.

[*Journal of Political Economy*, 1974, vol. 82, no. 6]

© 1974 by The University of Chicago. All rights reserved.

One reason for the absence of empirical research on this question is perhaps the fact that the usual procedures for testing whether a regression relationship is constant over time in the statistical sense, namely, the use of the *F*-test (see Chow 1960), or dummy variables, require prior knowledge of the point in time when the change in the estimated coefficients is suspected. Except for obvious dramatic changes, such as war, etc., this a priori knowledge may not be easily available.

The question of the stability of the demand for money is, however, of crucial importance, as the existence of this stable relationship is necessary for the effectiveness of monetary policy. In order to predict the effect of changes in the supply of money on variables such as income, prices, and the rate of interest, with any degree of certainty, one must be reasonably confident that the demand function itself does not shift. Indeed, the entire debate between the monetarists and Keynesians regarding the role of money in determining economic activity and the relative importance of monetary versus fiscal policy can be considered in the framework of the relative stability over time of the demand-for-money and expenditure functions.[2]

The purpose of this paper is to determine whether a stable demand-for-money function existed for the United States over the period 1901–65. This is done by utilizing a technique developed by Brown and Durbin (1968) that is designed to detect shifts in a regression relationship. This approach, as opposed to others mentioned above, does not require prior knowledge of the shifts but rather tests for the presence of such occurrences over the period of study. The United States was studied for two reasons. First, there exists ample evidence on the appropriate specification of the demand-for-money function (see Laidler 1966, 1971), and this allows one to abstract from the question of choosing appropriate explanatory variables and therefore limits the number of alternative functions to be tested. Second, the United States is considered to have undergone considerable structural change over the period, and previous studies have suggested that the demand-for-money function has shifted slowly over time (see Laidler 1971).

Although this paper is concerned solely with the analysis of stability, an important point to note is that the three outstanding issues referred to by Johnson (1962) are not independent of each other. As Laidler (1969) suggests, one should choose a definition of money for which the demand function is most stable when the same arguments are included. Therefore, in determining whether a given function is stable with respect to a number of theoretically appropriate variables, one is able simultaneously to determine the appropriate definition of the money aggregate.[3]

---

[2] For a detailed discussion of this, see Laidler (1971).

[3] Namely, whether it should be "narrow" or "broad." The former comprises currency and demand deposits, and the latter includes time deposits as well.

The outline of the paper is as follows. In Section II alternative specifications of the demand-for-money function for the United States are discussed. Section III describes the basic test of stability that is to be used. The results are discussed in Section IV, and their implications for the efficacy of monetary policy are considered in the concluding section. The data sources and definitions are presented in the Appendix.

II. Specification of the Demand-for-Money Function

There appears to be broad agreement in the literature on the appropriate specification of a demand-for-money function for the United States. In general terms, the demand for real money balances[4] in per capita terms is specified as a positive function of real per capita income (current or "permanent")[5] and a negative function of the opportunity cost of holding money. This opportunity cost is the yield on alternative assets and is therefore represented by the rate of interest.

In previous studies on the demand for money in the United States there does not emerge any clear-cut evidence in favor of using the level of current or permanent income[6] or the long-term or short-term rate of interest as the appropriate explanatory variables. Both definitions of income and the rate of interest will therefore be used in the analysis.

Formally the demand-for-money function in log-linear terms can be written as

$$\log M_t = a_0 + a_1 \log R_t + a_2 \log Y_t + u_t, \tag{1}$$

where M = real per capita money balances, narrow ($M1$) or broad ($M2$); R = rate of interest, short-term (SR) or long-term (LR); Y = real per capita income, current (Y) or permanent (Y^P); u_t is an error term; and a_1 and a_2 are the elasticities of money with respect to the rate of interest and income and are expected to carry the following signs: $a_1 < 0$; $a_2 > 0$.

It is evident from a previous study by Laidler (1971) over the same time period as this one that both real per capita money balances and real per capita income have followed pronounced trends in the United States. It is therefore more appropriate to specify the function in terms of

---

[4] Whether narrow ($M1$) or broad ($M2$).

[5] Permanent income is generated by the following process:

$$Y^P(t) = \lambda \int_{-\infty}^{T} e^{(\lambda - g)(-T+t)} Y(t)\, dt,$$

where Y^P is permanent income and Y is current income; λ is the coefficient of adjustment; and g is a growth factor.

[6] However, see Chow (1966).

first differences of the variables:[7]

$$\Delta \log M_t = b_0 + b_1 \Delta \log R_t + b_2 \Delta \log Y_t + w_t, \qquad (1a)$$

where the variables are defined as before. Delta is a difference operator, $\Delta \log M_t = \log M_t - \log M_{t-1}$; w_t is an error term with the property $w_t \sim \text{NID}(0, \sigma_w^2)$.

The use of two definitions of the money stock ($M1$ and $M2$) as alternative dependent variables and two definitions each of the interest rate (SR and LR) and income variables (Y and Y^P) yield eight specifications of the demand-for-money equation. Each of these will be tested to determine whether they were statistically stable over the period 1901–65.

III. Test of Stability

In general, statistical tests made to determine whether a regression relationship is stable over the entire period of observation have been limited either to using dummy variables when the change in the relationship is suspected, or to splitting the sample period at that point and performing the exact F-test devised by Chow (1960).[8] As mentioned in the Introduction, both these procedures require a priori knowledge of the point in time when the function shifts, and this knowledge may not be readily available. Alternative tests that do not require this a priori information have recently been devised by Goldfeld and Quandt (1972) and Fair and Jaffee (1972). These tests, based on calculating log likelihood ratios, are in general difficult to apply. A much simpler method of determining whether a regression relationship is constant over a period of time, based on an examination of the residuals of the relationship, has been developed by Brown and Durbin (1968).[9] This procedure will be used to test the stability of each of the eight specifications and is described in this section.

Basically, the Brown-Durbin method requires the calculation of one-period prediction residuals, which are obtained by applying the regression computed with $r - 1$ observations to predict the rth observation using k explanatory variables (including the constant). The method utilizes a statistic Sr which equals the ratio of the sum of squares of one-period prediction residuals from the $k + 1$ period to the rth period, to the sum of squares of one-period prediction errors from the $k + 1$ to the Tth period, T being the sample size. The null hypothesis that the regression relationship is constant over time implies that the expectation of the statistic Sr, $E(Sr)$, will lie along its mean-value line.[10]

---

[7] This is equivalent to introducing a linear time trend into (1). The constant term b_0 in (1a) can be viewed as the coefficient of the time trend in the original formulation.

[8] An application of this F-test in the context of the demand for money has been carried out by Courchene and Shapiro (1964).

[9] Several tests are proposed by Brown and Durbin (1968). The one used here appears to be the simplest to apply.

Consider the standard regression model:

$$Y_t = X_t'\beta_t + \varepsilon_t \qquad t = 1, \ldots, T, \tag{2}$$

where Y_t is a vector of observations on the dependent variable, x_t is a column vector of observations on the k regressors, β_t is a vector of regression coefficients (the subscript t implying that the β's may not be constant over time), and $\varepsilon_1, \ldots, \varepsilon_t$ are normal independent variables with zero means and variances $\sigma_1^2, \ldots, \sigma_t^2$. The first element in each x_t is one, and the remaining are nonstochastic.[11]

The hypothesis to be tested is

$$\beta_1 = \beta_2 = \cdots = \beta_t = \beta.$$

Let $\hat{\beta}_r$ be the least-squares estimate of β from the sample of the first r observations, and let

$$w_r = \frac{y_r - x_r'\hat{\beta}_{r-1}}{\sqrt{1 + x_r'(X_{r-1}'X_{r-1})^{-1}x_r}}, \qquad r = k + 1, \ldots, T, \tag{3}$$

where $X_{r-1}' = [x_1, \ldots, x_{r-1}]$.

It can be shown that these transformed residuals, $w_{k+1}, \ldots, w_t$, are normal independent variables with zero means and variances, $\sigma_{wk+1}^2, \ldots, \sigma_{wt}^2$, if $\varepsilon_1, \ldots, \varepsilon_t$ are normal independent. Equation (3) is a generalization of the Helmert Orthogonal Transformation (see Brown and Durbin 1968).

The variables w_r can be obtained without repeated matrix inversion by the relations

$$\hat{\beta}_r = \hat{\beta}_{r-1} + (X_r'X_r)^{-1}x_r'(y_r - x_r'\hat{\beta}_{r-1}) \tag{4}$$

and

$$(X_r'X_r)^{-1} = (X_{r-1}'X_{r-1})^{-1} - \frac{(X_{r-1}'X_{r-1})^{-1}x_r x_r'(X_{r-1}'X_{r-1})^{-1}}{1 + x_r'(X_{r-1}'X_{r-1})^{-1}x_r}. \tag{5}$$

If β_t is constant up to time $t = t_0$ and different from then on, the w_r will have mean zero up to t_0 and nonzero mean from then on. Brown and Durbin suggest plotting the following variable against time:

$$Sr = \left(\sum_{k+1}^{r} w_t^2\right) \Bigg/ \left(\sum_{k+1}^{T} w_t^2\right) \qquad r = k + 1, \ldots, T, \tag{6}$$

that is, the value of the cumulative sum of w_r^2 divided by the cumulated sum of w_t^2 over the entire period. The value of Sr will thus lie between zero and one: $Sr = 0$ if $r < k + 1$; $Sr = 1$ if $r = T$. The expectation of Sr is $E(Sr) = (r - k)/(T - k)$, and on the null hypothesis that the β's are constant, the plot of Sr should lie along this mean-value line.

---

[10] The sample Sr's will, of course, vary about the mean-value line.

[11] Therefore excluding lagged dependent variables and autoregressive schemes.

Significance tests (for a two-sided test) can be performed by drawing a pair of lines $Sr \pm c_0 + (r - k)/(T - k)$ parallel to the mean-value line. If the sample path of Sr crosses either of these significance lines, the null hypothesis of the constancy of the regression coefficients can be rejected. The statistics c_0 are distributed as Pyke's modified Kolmogorov-Smirnov statistic,[12] and values for various significance levels are provided in Durbin (1969, table 1), entering the table at $m = \frac{1}{2}(T - k)$ when $T - k$ is even and interpolating linearly between $m = \frac{1}{2}(T - k) - \frac{1}{2}$ and $m = \frac{1}{2}(T - k) + \frac{1}{2}$ when $T - k$ is odd.

This test of stability is approximate and tends to be somewhat optimistic. However, it is expected that the approximation is good when T is large.

A forthcoming paper by J. Durbin and P. B. Kenny will contain theoretical proofs and discuss these tests further.

The operational procedure can thus be summarized by the following steps:[13] (1) calculate Sr and plot against time; (2) draw the mean-value line, $E(Sr)$; and (3) for a two-sided test, draw a pair of lines $\pm c_\alpha + (r - k)/(T - k)$ parallel to the mean-value line, where α is the required significance level.

The results of each of the eight specifications of Section II, both the ordinary least-squares results for the period 1901–65 and the stability tests, are shown in the following section.

IV. Results

The results from estimating the eight alternative specifications by ordinary least squares on an annual basis, 1901–65, are shown in table 1. Data definitions and sources are in the Appendix. All estimated elasticities in the regressions have the correct signs and are significantly different from zero at the 5 percent level. The current income elasticities are somewhat on the low side for both definitions of money, while the permanent income elasticities are similar to those obtained by Laidler (1966) and others. The long-term rate of interest appears to exert a stronger effect on both narrow and broad money balances than does the short-term rate.[14]

The corrected R^2's are low in all the equations, but they are nevertheless significantly different from zero in each case. The most important

---

[12] For the exact definition of c_0, see Brown and Durbin (1968) and Durbin (1969).

[13] The program used to derive w_r and Sr, utilizing relationships (4) and (5), has been written at the IMF and is available for the Burroughs 5700 system. I am indebted to my colleague, R. Roadman, for programming assistance.

[14] See Khan (1973) for a similar result for velocity.

TABLE 1

ORDINARY LEAST-SQUARES ESTIMATES, 1901–65

	Constant	$\Delta \log Y_t$	$\Delta \log Y_t^P$	$\Delta \log SR_t$	$\Delta \log LR_t$	R^{2*} (F)	SE	D.-W.†
$\Delta \log M1$............	0.009 (1.47)	0.332 (3.69)	...	−0.085 (2.86)	...	0.208 (9.39)	0.049	1.60
	0.012 (2.03)	0.208 (2.49)	...	...	−0.303 (4.61)	0.332 (16.92)	0.046	1.66
	−0.003 (0.48)	...	1.090 (5.73)	−0.122 (4.38)	...	0.368 (19.61)	0.045	1.83
	0.005 (0.77)	...	0.653 (3.69)	...	−0.298 (4.82)	0.398 (22.13)	0.044	1.76
$\Delta \log M2$............	0.017 (2.99)	0.279 (3.46)	...	−0.070 (2.65)	...	0.183 (8.18)	0.045	1.63
	0.019 (3.63)	0.175 (2.32)	...	...	−0.256 (4.30)	0.300 (14.71)	0.041	1.50
	0.005 (0.92)	...	1.021 (6.24)	−0.107 (4.47)	...	0.401 (22.40)	0.038	1.59
	0.012 (2.17)	...	0.645 (4.61)	...	−0.246 (4.53)	0.405 (22.80)	0.038	1.69

NOTE.—t-values are given in parentheses below coefficients.
* Coefficient of determination corrected for degrees of freedom.
† Durbin-Watson test statistic.

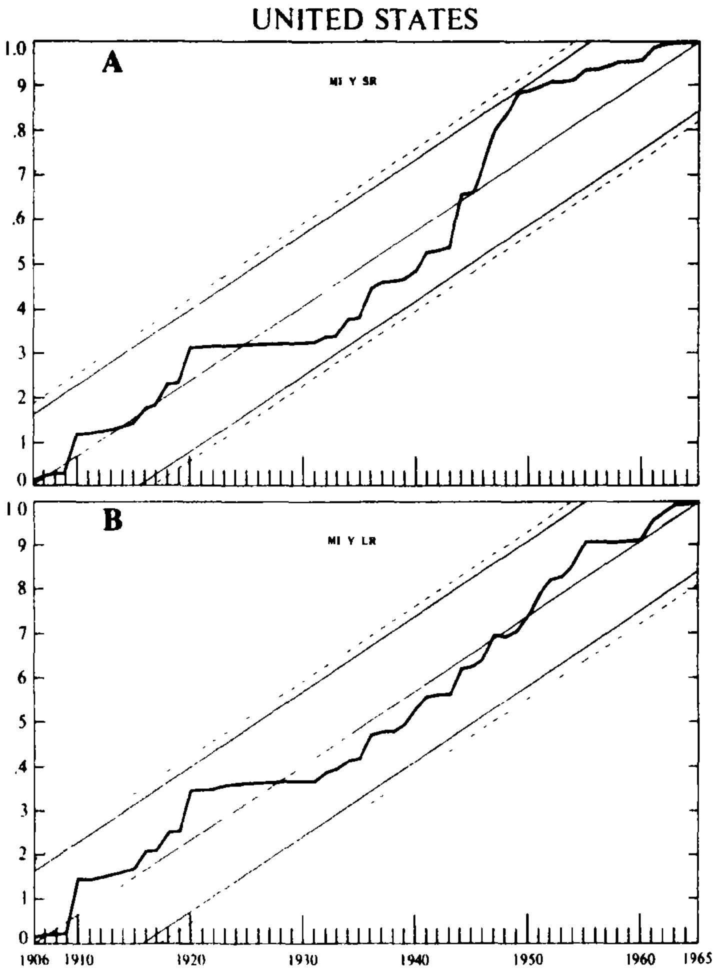


Fig. 1

result from the point of view of the stability test is that the null hypothesis of the serial independence of the error term is accepted at the 5 percent level (see Section III).

The sample plots of the statistic *Sr* for each of the specifications of the demand for money are shown in eight diagrams, in figures 1–4. Each diagram is identified by the dependent variable and the two relevant

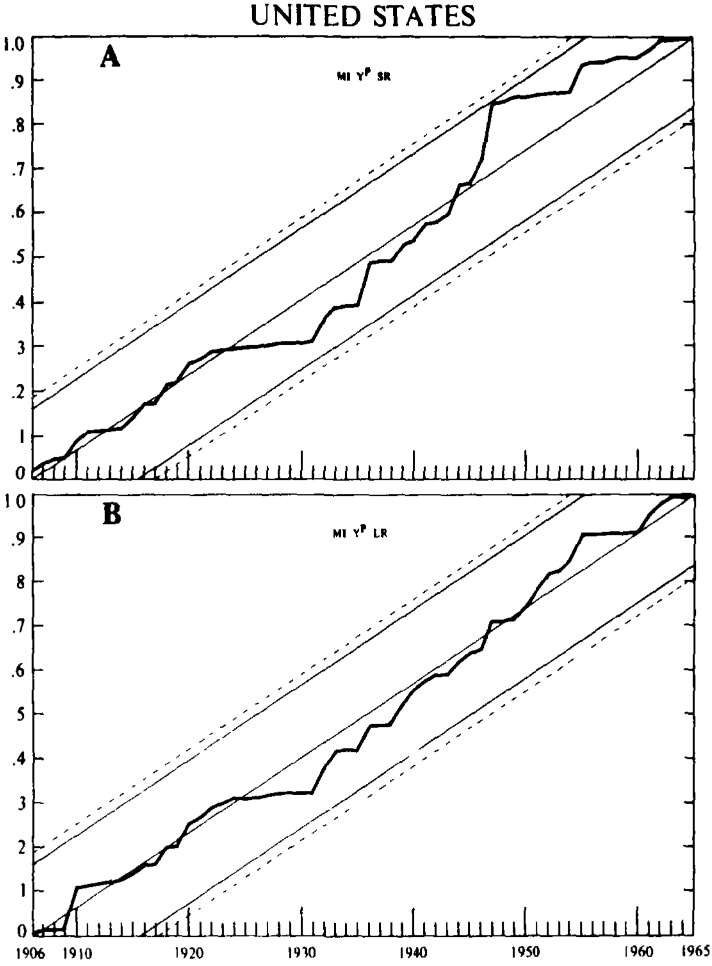


Fig. 2

explanatory variables. The period over which the test was made is 1906–65.[15] On each of the diagrams is drawn the mean-value line, $E(Sr) = (r - k)/(T - k)$, and two sets of confidence lines parallel to the mean-value line. The set of dotted lines corresponds to a significance

---

[15] The process should start from 1905, i.e., $k + 1$, but in all specifications the 1905 value of Sr was zero up to the fourth decimal place.

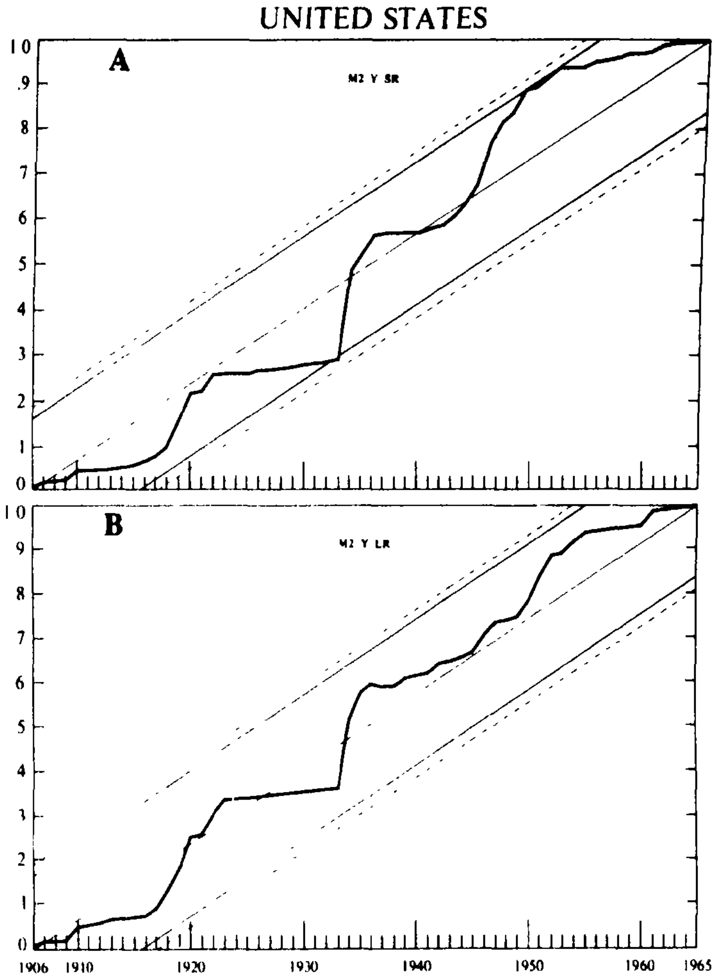


Fig. 3

level of 5 percent, while the other set corresponds to a significance level of 10 percent. The values of these significance lines were obtained from Durbin (1969, table 1).

In the diagrams for narrow money balances, figures 1–2, it can be observed that the sample paths of Sr do not travel outside the 5 percent confidence band. However, in those specifications using the short-term rate of interest as an explanatory variable, that is, figures 1*A* and 2*A*,

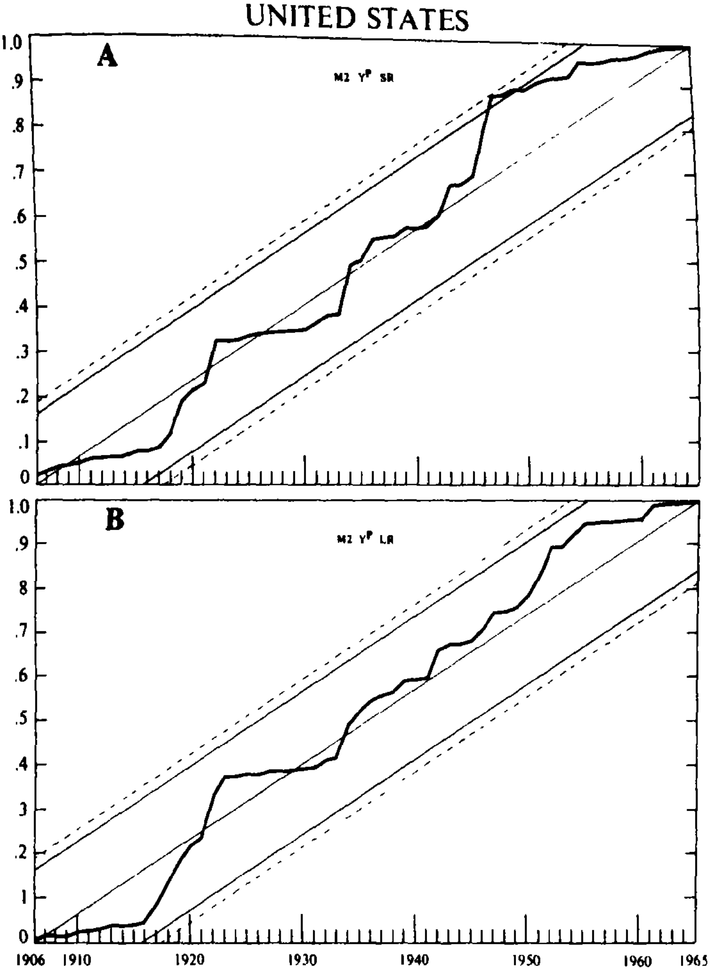


Fig. 4

the *Sr* appears to cross the upper 10 percent significance line at about 1948. In the context of the test of stability, this implies a structural shift in the relationship at that point.[16] When the long-term rate of interest is used, however, there is no evidence of structural change.

A similar picture emerges from the diagrams for broad-money-balances

---

[16] This is not strictly true, as the shift could have taken place earlier and only became significant at this point in time. I am indebted to Professor R. E. Quandt for this point.

TABLE 2

Variables	F-ratio
$M1, Y, SR$	1.76
$M1, Y^P, SR$	1.20
$M2, Y, SR$	2.15
$M2, Y^P, SR$	1.31

specifications—figures 3–4—namely, that the *Sr*'s calculated do not cross either of the 5 percent significance lines. When the short-term rate is used as an explanatory variable, figures 3*A* and 4*A*, the *Sr*'s cut the upper 10 percent significance line at around 1948. When the long-term rate of interest is used, the regression relationship once again appears to be constant, since the *Sr*'s stay within the bands.

The most interesting feature of these results for both definitions of money is that a structural shift is implied at the same point of time, namely, 1948, when the short-term rate of interest is used. It does not appear to matter whether current real per capita income or permanent real per capita income is used as the second explanatory variable. There does not seem to be any a priori reason why the shift should occur in 1948, and other tests have generally considered the 1930s or the war years 1917 19 and 1942–45 as periods when shifts in the demand-for-money function occurred.[17]

The F-test devised by Chow (1960) was applied in the case of the four specifications using the short-term rate of interest as an independent variable. The sample was thus split into two periods, 1901–48 and 1949–65, and the sum of squared residuals was calculated. If we keep in mind the qualification noted previously regarding the precise timing of the shift, the results should prove to be informative. The necessary F-statistics calculated are shown in table 2.

The critical values for F with 3.59 degrees of freedom at the 5 percent and 10 percent levels are 2.25 and 2.17, respectively. The results of the Brown-Durbin test do not appear to be confirmed by the F-test, but it can be argued that the F-test requires the precise point in time when the sample is divided, whereas the Brown-Durbin test gives only the approximate region.[18] Viewed in this way, the results are not necessarily in conflict.

---

[17] With the exception of Courchene and Shapiro (1964), who applied an F-test after splitting their sample into two periods, 1900–1929, and 1930–58, the other tests have been based on an examination of parameters before and after the breakpoint (see Laidler 1966, 1971). The Courchene-Shapiro tests cannot strictly be compared with those here because their equations have serial correlation and the F-test is inapplicable in such situations.

[18] This region could be narrowed by doing the analysis in the reverse direction or performing the test around this region.

Two further interesting results emerge from the diagrams. First, in each specification there is a common movement in *Sr* in about the 1920–30 period. Though it does not represent a significant shift in the relationship, the result is nevertheless evident from the diagrams. The 1930s are generally regarded as a period when the demand-for-money function was unstable, but the results here indicate that although sharp erratic movements in *Sr* do take place, the function does not shift significantly. Second, periods that are generally omitted from the sample, namely, the war years 1917–19 and 1942–45, were included to determine whether the function shifted during these abnormal years.[19] In no case does there appear to have been a significant shift.

V. Conclusion

Laidler's (1971) view that the evidence of the existence of a stable relationship between the demand for real balances and a few variables is "overwhelming" is supported by the results in this paper for the United States for the period 1901–65. This is especially true if the "few" variables include a long-term rate of interest rather than a short-term rate. It does not appear to matter whether current real per capita income or the longer-run concept of permanent real per capita income is used as the other explanatory variable. Stability, defined in the sense of constancy of parameters over time, is apparent for those demand-for-money functions including the long-term rate of interest. This conflicts with Laidler's (1971) results, which indicate that the most stable demand-for-money function is one that includes the short-term rate of interest and permanent income as explanatory variables.

The implications of the results for the relative-stability debate, namely, the stability of the demand-for-money function as against the expenditure function, are not clear-cut, since only one side of the argument has been investigated. Although it was established that stability is apparent for certain demand-for-money functions, a proper test would have been also to examine the stability of the expenditure function over the same time period. This exercise is, however, outside the scope of this paper. Nevertheless, it has been demonstrated that a necessary condition for monetary policy to have an effect on macroeconomic variables is met over a long time period in the United States.

Since both narrow and broad money equations were apparently stable, a choice cannot be made on the appropriate definition of the money stock by appealing to the stability criterion, as has been suggested by Laidler (1969). The question of definition must still remain open.

---

[19] This conforms to Friedman's (1959) approach of including these years in the sample period.

Appendix

Data Definitions and Sources

$M1$ = currency plus demand deposits at commercial banks (Friedman and Schwartz 1970).
$M2$ = $M1$ plus time deposits (Friedman and Schwartz 1970).
Y = net national product (NNP) at 1929 prices (Kuznets 1961; series extended to 1965 from unpublished NBER series).
Y_p = permanent net national product (unpublished NBER series).
LR = Moody's Corporate Aaa bond yield (U.S. Bureau of the Census 1960; 1966).
SR = 4–6-month commercial paper rate (U.S. Bureau of the Census 1960; 1966).
Prices = NNP deflator, 1929 = 1.00 (Kuznets 1961; series extended to 1965 from unpublished NBER series).

Sources of population data: U.S. Bureau of the Census 1960; 1966.

References

Brown, R. L., and Durbin, J. "Methods of Investigating Whether a Regression Relationship Is Constant over Time." Paper presented at the European Statistical Meeting, Amsterdam, 1968.

Chow, G. "Tests of Equality between Two Sets of Coefficients in Two Linear Regressions." *Econometrica* 28 (July 1960): 591–605.

———. "On the Long-Run and Short-Run Demand for Money." *J.P.E.* 74, no. 2 (April 1966); 111–31.

Courchene, T. J., and Shapiro, H. T. "The Demand for Money: A Note from the Time Series." *J.P.E.* 72, no. 5 (October 1964): 498–503.

Durbin, J. "Tests for Serial Correlation in Regression Analysis Based on the Periodogram of Least-Squares Residuals." *Biometrika* 56 (March 1969): 1–15.

Fair, R. C., and Jaffee, D. M. "Methods of Estimation for Markets in Disequilibrium." *Econometrica* 40 (May 1972): 497–514.

Friedman, M. "The Quantity Theory of Money—a Restatement." In *Studies in the Quantity Theory of Money,* edited by M. Friedman. Chicago: Univ. Chicago Press, 1956.

———. "The Demand for Money: Some Theoretical and Empirical Results." *J.P.E.* 67, no. 4 (August 1959): 327–51.

Friedman, M., and Schwartz, A. *Monetary Statistics of the United States.* New York: Nat. Bur. Econ. Res., 1970.

Goldfeld, S. M., and Quandt, R. E. *Nonlinear Methods in Econometrics.* Amsterdam: North-Holland, 1972.

Johnson, H. G. "Monetary Theory and Policy." *A.E.R.* 52 (June 1962): 335–84.

Khan, M. S. "A Note on the Secular Behavior of Velocity within the Context of the Inventory-theoretic Model of Demand for Money." *Manchester School* 41 (June 1973): 207–13.

Kuznets, S. *Capital in the American Economy.* New York: Nat. Bur. Econ. Res., 1961.

Laidler, D. E. W. "The Rate of Interest and the Demand for Money—Some Empirical Evidence." *J.P.E.* 74, no. 6 (December 1966): 543–55.

———. "The Definition of Money—Theoretical and Empirical Problems." *J. Money, Credit, and Banking* 1 (August 1969): 508–25.

———. "A Survey of Some Current Problems." In *Monetary Theory and Monetary Policy in the 1970's*, edited by G. Clayton, J. C. Gilbert, and R. Sedgewick. London: Oxford Univ. Press, 1971.

Meltzer, A. H. "The Demand for Money: The Evidence from the Time Series." *J.P.E.* 71 (June 1963): 219–46.

U.S., Bureau of the Census. *Historical Statistics of the United States.* Washington: Government Printing Office, 1960.

———. *Statistical Abstract of the United States.* Washington: Government Printing Office, 1966.

Estimating the Long-Run Demand for Money from Time-Series Data

---

Rodney L. Jacobs

Stanford University

A simple theoretical model is developed to illustrate that three aggregation procedures used in the estimation of long-run money demand functions from time-series data—deflating the aggregate data by population and prices, deflating by prices only, or using nominal data undeflated by population or prices—are mathematically equivalent when the data are dominated by time trend. Differences in regressions based on the three aggregation procedures have nothing to do with the degree of homogeneity in population and prices, as is often claimed, but merely reflect common time trends in the data. The model is seen to provide good agreement with data from three countries.

Introduction

In his restatement of the quantity theory, Friedman (1956) develops the two central themes of the "new monetarists." First, the quantity theory is a theory of the demand for real money balances rather than a theory of the price level. Second, the quantity theory rests on the empirical generalization that the demand for money, defined over some narrow group of financial assets such as currency, demand deposits, and time deposits, is stable over time. The stable demand function implies that any change in the narrow group of financial assets (the money supply) will have a direct impact on expenditure decisions.

Since the Friedman article, there has been a proliferation of empirical work on money demand functions. The usual approach to this empirical work has been to postulate an individual money demand function which depends on the opportunity cost of holding real money balances and on some measure of income or wealth. To estimate the money demand func-

The author wishes to thank the National Science Foundation for financial support and an anonymous referee for indicating a serious error in an earlier version of the paper.
[*Journal of Political Economy*, 1974, vol. 82, no. 6]
© 1974 by The University of Chicago. All rights reserved.

tion from aggregate time-series data, it is necessary to aggregate the individual demand functions over all individuals. Three different aggregation procedures have been employed in the literature. Some investigators have deflated the aggregate monetary and income[1] data by population and prices and have estimated the demand function in real per capita terms. Other investigators have only deflated by prices and have estimated a demand function in aggregate real terms. Finally, some investigators have ignored the distinction between real and nominal money balances and have estimated the demand function from aggregate time-series data undeflated by population or prices. In all instances, the coefficient of the logarithm of the income variable is taken to be the income elasticity of money demanded. The usual justification for choosing one of the latter two aggregation procedures is that the per capita demand for real money balances is not homogeneous of degree zero in population or prices. For instance, Chow (1966) states that the money demand function should be estimated in real terms if money illusion is absent and in nominal terms if money illusion is assumed.

It is the purpose of this paper to show that in much of the empirical work on money demand functions, differences in the three aggregation procedures have had nothing to do with the degree of homogeneity in prices and population, as has often been claimed, but have merely reflected common time trends in the data. There appears to be no justification for using the latter two aggregation procedures, deflating by prices or using undeflated data, since they provide no new information and can be quite misleading if not interpreted properly.

A Theoretical Perspective

In this section we will develop a theoretical perspective from which to view the various aggregation procedures employed in the estimation of long-run money demand functions from time-series data. We begin by defining

$$\begin{aligned} Z_{1i} &= \ln (M/NP)_i, & X_{1i} &= \ln (Y/NP)_i, \\ Z_{2i} &= \ln (M/P)_i, & X_{2i} &= \ln (Y/P)_i, \\ Z_{3i} &= \ln (M)_i, & X_{3i} &= \ln (Y)_i, \\ V_i &= \ln (N)_i, & W_i &= \ln (P)_i, \end{aligned}$$

where M is the aggregate money stock, N is the population, P is the price level, Y is aggregate nominal income, and the subscript i denotes the ith observation of a time-series containing T observations. Since most time-series data on population, prices, and income are dominated by a trend factor, we will assume that these variables are pure trend. This assumption implies that $V_t = V_0 + \dot{N}i$, $W_t = W_0 + \dot{P}i$, $X_{1t} = X_{10} + \dot{y}i$, $X_{2t} = X_{20} + (\dot{N} + \dot{y})i$, and $X_{3t} = X_{30} + (\dot{P} + \dot{N} + \dot{y})i$, where y is per capita real income, Y/NP, and $\dot{S}$ denotes $(1/S)(dS/dt)$ for any variable S. We further assume that the rate of interest, r, cannot be approximated by a pure trend or a constant.

---

[1] Those readers who prefer may substitute wealth for income in the remainder of the paper.

We now consider a commonly used money demand function formulated in real per capita variables which is homogeneous of degree zero in prices and population and is of the form

$$Z_{1t} = \alpha_1 + \delta_1 X_{1t} + \gamma_1 \ln (r)_t + U_{1t}, \tag{1}$$

where δ_1 is the income elasticity of real money demanded and U_{1t} is a stochastic variable assumed to be independently and identically distributed with mean zero and finite variance. If an investigator were to regress ln (M/NP) on ln (Y/NP) and ln (r), he would obtain $E(\hat{\delta}_1) = \delta_1$ as an estimate of the income elasticity of money demanded.

In order to consider the regression based on aggregate real variables, we expand equation (1) to obtain

$$Z_{2t} = \alpha_1 + \delta_1 X_{2t} + (1 - \delta_1)V_t + \gamma_1 \ln (r)_t + U_{1t}.$$

Since $X_{2t} = X_{20} + (\dot{N} + \dot{y})i$ and $V_t = V_0 + \dot{N}i$, we can solve for V_t in terms of X_{2t} to obtain

$$V_t = \left(\frac{\dot{N}}{\dot{N} + \dot{y}}\right) X_{2t} + \left(V_0 - \frac{\dot{N}}{\dot{N} + \dot{y}} X_{20}\right).$$

Substituting this expression for V_t into the equation for Z_{2t} and rearranging terms, we obtain[2]

$$Z_{2t} = \alpha_2 + \delta_2 X_{2t} + \gamma_2 \ln (r)_t + U_{2t}, \tag{2}$$

where

$$\alpha_2 = \alpha_1 + (1 - \delta_1)\left(V_0 - \frac{\dot{N}}{\dot{N} + \dot{y}} X_{20}\right), \tag{3}$$

$$\delta_2 = (\dot{N} + \delta_1 \dot{y})/(\dot{N} + \dot{y}), \tag{4}$$

$$\gamma_2 = \gamma_1, \tag{5}$$

and

$$U_{2t} = U_{1t}. \tag{6}$$

---

[2] The equations could also be viewed as a case of left-out variables where $Z_{2t} = \alpha_1 + \delta_1 X_{2t} + \gamma_1 \ln (r)_t + (1 - \delta_1)V_t + U_{1t}$ is truncated to give $Z_{2t} = \alpha_1 + \delta_1 X_{2t} + \gamma_1 \ln (r)_t + U_{1t}$ by *omitting V_t from the regression*. However, V_t and X_{2t} are collinear so the expanded version of the demand equation could not be estimated. Making use of the collinearity by substituting for V_t provides a simple solution compared with the truncated-variables algorithm.

The real per capita variables formulation of the demand function can thus be reduced to the aggregate real variables formulation. If an investigator were to regress ln (M/P) on ln (Y/P) and ln (r), he would obtain $E(\hat{\delta}_2) = (\dot{N} + \hat{\delta}_1\dot{y})/(\dot{N} + \dot{y})$ as an estimate of the income elasticity of money demanded. If we assume that the correlation between ln (r) and the other independent variables is negligible, it can also be shown that

$$\operatorname{var}(\hat{\alpha}_2) = \frac{\dot{y}^2}{(\dot{N}+\dot{y})^2} \frac{TX_{20}^2 + 2(\dot{y}+\dot{N})X_{20}\sum i + (\dot{y}+\dot{N})^2\sum i^2}{TX_{10}^2 + 2\dot{y}X_{10}\sum i + \dot{y}^2\sum i^2} \operatorname{var}(\hat{\alpha}_1) \tag{7}$$

and

$$\operatorname{var}(\hat{\delta}_2) = \frac{\dot{y}^2}{(\dot{N} + \dot{y})^2} \operatorname{var}(\hat{\delta}_1). \tag{8}$$

Thus, if $\hat{\delta}_2$ is interpreted as the income elasticity of money demanded, we see that it is biased toward unity compared with $\hat{\delta}_1$ and that its variance can be considerably smaller than the variance of $\hat{\delta}_1$.

In order to consider the regression based on aggregate nominal variables, we expand equation (1) to obtain

$$Z_{3t} = \alpha_1 + \delta_1 X_{3t} + (1 - \delta_1)(V_t + W_t) + \gamma_1 \ln (r)_t + U_{1t}.$$

Substituting the linear expressions for X_{3t}, V_t, and W_t and performing some algebraic manipulation, we obtain

$$Z_{3t} = \alpha_3 + \delta_3 X_{3t} + \gamma_3 \ln (r)_t + U_{3t}, \tag{9}$$

where

$$\alpha_3 = \alpha_1 + (1 - \delta_1)\left(V_0 + W_0 - \frac{\dot{N} + \dot{P}}{\dot{N} + \dot{P} + \dot{y}} X_{30}\right), \tag{10}$$

$$\delta_3 = (\dot{N} + \dot{P} + \delta_1\dot{y})/(\dot{N} + \dot{P} + \dot{y}), \tag{11}$$

$$\gamma_3 = \gamma_1, \tag{12}$$

and

$$U_{3t} = U_{1t}. \tag{13}$$

Thus, an investigator who regresses ln (M) on ln (Y) and ln (r) would obtain

$$E(\hat{\delta}_3) = (\dot{N} + \dot{P} + \hat{\delta}_1\dot{y})/\dot{B},$$

$$\operatorname{var}(\hat{\alpha}_3) = \frac{\dot{y}^2}{\dot{B}^2} \frac{TX_{30}^2 + 2\dot{B}X_{30}\sum i + \dot{B}^2\sum i^2}{TX_{10}^2 + 2\dot{y}X_{10}\sum i + \dot{y}^2\sum i^2} \operatorname{var}(\hat{\alpha}_1), \tag{14}$$

and

$$\operatorname{var}(\hat{\delta}_3) = \frac{\dot{y}^2}{\dot{B}^2} \operatorname{var}(\hat{\delta}_1), \tag{15}$$

where $\dot{B} = \dot{N} + \dot{P} + \dot{y}$. If $\hat{\delta}_3$ is interpreted as the income elasticity of money demanded, we see that it is biased toward unity compared with $\hat{\delta}_1$ and has a considerably smaller variance.

The original demand function of equation (1) was assumed to be homogeneous of degree zero in population and prices, and δ_1 was taken to be the income elasticity of money demanded. But equation (1) is merely a special case of the demand function

$$Z_{1t} = \alpha^* + \delta^* X_{1t} + \gamma_1 \ln (r)_t + \psi V_t + \Theta W_t + U_{1t}, \qquad (16)$$

where ψ and Θ are the degree of homogeneity in population and prices of the per capita demand for real money balances and δ^* is the real per capita income elasticity of money demanded. Since X_{1t}, V_t, and W_t are collinear under the assumptions of our model, the above equation cannot be estimated directly. By combining the collinear terms, equation (16) can be reduced to equation (1):

$$Z_{1t} = \alpha_1 + \delta_1 X_{1t} + \gamma_1 \ln (r)_t + U_{1t},$$

where

$$\alpha_1 = \alpha^* + \psi\left(V_0 + X_{10}\frac{\dot{N}}{\dot{y}}\right) + \Theta\left(W_0 - X_{10}\frac{\dot{P}}{\dot{y}}\right) \qquad (17)$$

and

$$\delta_1 = \delta^* + \psi\frac{\dot{N}}{\dot{y}} + \Theta\frac{\dot{P}}{\dot{y}}. \qquad (18)$$

Thus, δ_1 cannot be interpreted as the income elasticity of money demanded as originally assumed but is a combination of the income elasticity and the degree of homogeneity in prices and population. Furthermore, although δ^* may be indeterminate because of collinearity, an investigator using any of the three standard formulations would conclude that the apparent income elasticity, δ_k, was well determined with a small standard error.

When the data are not pure trend, equation (16) represents a more general form of the money demand function and can be estimated directly. The functional relationship between the parameters of equation (16) and equation (1) or between the parameters of equation (1) and equations (2) and (9) would then be different from the formulae derived above. Equation (1) would still be a legitimate equation to estimate whether empirical evidence or theoretical considerations indicated that ψ and Θ were zero. However, differences between equation (1) and equations (2) and (9) would still have nothing to do with deviations of ψ and Θ from zero, and there would still be no justification for using the latter two estimation equations. The pure trend model has been developed because the data are dominated by trend and because it allows a consistent interpretation of previous estimates of the demand for money.

Comparison of Theory and Data

In the preceding section we demonstrated that three commonly used formulations of a money demand function were mathematically equivalent when the time-series data are dominated by time trend. Differences between the three equations have nothing to do with a lack of homogeneity of degree zero in prices and population but merely reflect the common time trend of the variables. It was also seen that the coefficient of the income variable was biased toward unity and had a smaller variance for the aggregate real and aggregate nominal formulations compared with the real per capita formulation of the money demand equation. However, the coefficient of the income variable cannot be interpreted as the income elasticity of money demanded for any of the three common formulations. The equations were also shown to be equivalent to a more general model which was not homogeneous of degree zero in prices and population. In this section we will demonstrate that this simple theoretical model provides a framework for interpreting regressions based on monetary data for three countries.

First, we consider estimates by Laidler and Parkin (1971) of the demand for money in the United Kingdom. Laidler and Parkin contend that Fisher (1968) overstated the explanatory power of simple formulations of the demand for money, such as equation (9), in Britain because he failed to deflate the data by prices and population. They then estimate the real per capita version of the demand function given by equation (1) and conclude that the simple formulations, if properly deflated by prices and population, do not provide a good fit to the data. What they fail to realize is that their formulation based on per capita real variables and Fisher's formulation based on aggregate nominal variables could be equivalent if viewed in the framework discussed above and that no conclusions could be drawn from differences in the two formulations. To investigate this possibility, we will examine a number of regressions on the United Kingdom data.

The regressions are based on seasonally adjusted quarterly data from the second quarter of 1956 to the fourth quarter of 1967.[3] Our three standard versions of the long-run money demand function are presented in table 1 as regressions 1.1, 1.3, and 1.5. The regression based on nominal data yields a higher R^2 and a larger apparent income elasticity with

---

[3] The money supply data are taken from appendix A of Laidler and Parkin. Income, which is gross domestic product, and the price index, which is the implicit price deflator for gross domestic product, are obtained from *Economic Trends* (Central Statistical Office 1968*b*). The interest rate is the 3-month Treasury bill rate and is derived by averaging monthly values presented in the *Annual Abstract of Statistics* (Central Statistical Office 1968*a*). The quarterly population values were obtained by interpolating the annual values of total United Kingdom home population listed in the *Annual Abstract of Statistics* (Central Statistical Office 1968*a*)

TABLE 1

UNITED KINGDOM EQUATIONS BASED ON QUARTERLY DATA FROM 1956 TO 1967
EQUATIONS $Z_{kt} = \alpha_k + \delta_k X_{kt} + \gamma_k \ln (r)_t + U_{kt}$

EQUATION No.	VARIABLES	METHOD	COEFFICIENT (STANDARD ERROR IN PARENTHESIS)			R^2	SSR	$v \approx F(3, 44)$
			α_k	δ_k	γ_k			
1.1	$Z_{kt} = \ln (M/NP)_t$	Regression	4.168 (0.192)	0.207 (0.043)	0.047 (0.017)	.539	0.02128	. . .
1.2	$X_{kt} = \ln (Y/NP)_t$ $(k = 1)$	Constraint	4.171	0.206	0.047	. . .	0.02128	0.0008
1.3	$Z_{kt} = \ln (M/P)_t$	Regression	14.288 (0.742)	0.387 (0.034)	0.044 (0.017)	.913	0.02150	. . .
1.4	$X_{kt} = \ln (Y/P)_t$ $(k = 2)$	Constraint	14.204 (0.659)	0.388 (0.032)	0.047	. . .	0.02154	0.0209
1.5	$Z_{kt} = \ln (M)_t$	Regression	7.771 (0.428)	0.676 (0.019)	0.036 (0.018)	.976	0.02542	. . .
1.6	$X_{kt} = \ln (Y)_t$ $(k = 3)$	Constraint	7.809 (0.362)	0.674 (0.018)	0.047	. . .	0.02565	0.0766

smaller standard error than the regression based on real per capita data. For the last two regressions, theoretical estimates of the coefficients α_k and δ_k and their standard errors, computed from equations (3)–(15), are given in rows 1.4 and 1.6 of table 1. The growth rates $\dot{N}$, $\dot{P}$, and $\dot{y}$ and the constants X_{k0}, V_0, and W_0 used in the equations were obtained from regressions of the population, price, and income logarithms on time. The parameters α_1, δ_1, and γ_1 were approximated by their estimates taken from regression 1.1. The figures in table 1 suggest that the last two regressions are merely alternate ways of writing the first regression. This suggestion can be tested by the F-test of a linear hypothesis. The hypothesis is that α_k, δ_k, and γ_k are constrained to equal their theoretical values. If we assume that all parameters on the right-hand side of the equations take on their true values, then the variable

$$v = \frac{T - K}{q} \frac{SSR^+ - SSR}{SSR}$$

is distributed as $F(q, T - K)$, where T is the number of data points, K is the number of regression parameters, q is the number of constraints, SSR^+ is the constrained sum-squared residuals, and SSR is the unconstrained sum-squared residuals. Since we only have estimates of the right-hand side variables of equations (3)–(5) and (10)–(12), obtained from the same data set as the regressions of table 1, the distribution of v will only be approximated by the F distribution. The results of the F-tests are presented in the last column of rows 1.4 and 1.6 of table 1. The values of v indicate that there are no significant differences between the theoretical and regression equations.

We now determine whether regression 1.1 could have been derived from the more general money demand function given by equation (16). If the independent variables were pure trend, then equation (16) could not be estimated because of multicollinearity. However, the deviations from pure trend are sufficient to allow the estimate

$$\begin{aligned} Z_{1t} = \underset{(26.55)}{4.767} + \underset{(0.26)}{0.230 X_{1t}} + \underset{(0.02)}{0.047 \ln (r)_t} - \underset{(1.55)}{0.039 V_t} - \underset{(0.25)}{0.011 W_t}, \quad (19) \\ R^2 = .539, \qquad SSR = 0.02127. \end{aligned}$$

Row 1.2 of table 1 presents an estimate of regression 1.1 derived from regression (19) and equations (17) and (18). The F-test in the last column indicates that there is no significant difference between the two expressions and that regression 1.1 can be derived from regression (19). The large standard errors shown in parentheses for equation (19) suggest a significant degree of collinearity, so that ψ and Θ could be constrained to arbitrary values with the differences being absorbed in the parameters α^* and δ^*. This possibility was tested by a series of F-tests in which ψ

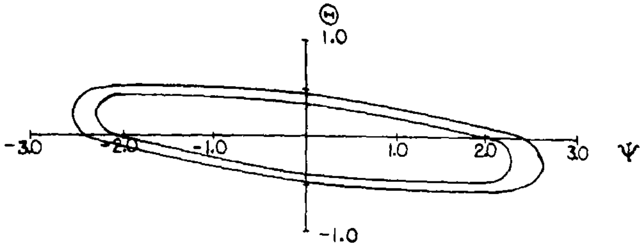


FIG. 1

and Θ of equation (16) were constrained to arbitrary values while the remaining parameters were fit. The series of F-tests gives the 95 and 99 percent confidence regions on ψ and Θ, which are presented in figure 1. Although the apparent income elasticity of the three regressions of table 1 has a small standard error, the range of ψ and Θ in the 95 percent confidence region gives a range of δ^* from -0.4 to 0.8 in the constrained regressions of equation (16).

The United Kingdom data were not used primarily to criticize a minor part of the Laidler and Parkin paper but because the data dramatically illustrate our two main points concerning estimates of the demand for money from data dominated by time trend. First, differences between the three commonly used aggregation procedures have nothing to do with the degree of homogeneity in prices and population but merely reflect common time trends of the data. Second, unless a priori restrictions can be placed on the degree of homogeneity with respect to prices and population, the income elasticity will be poorly determined. The regressions of table 1, however, should not be considered realistic estimates of the long-run demand for money in Great Britain. The low value of R^2 for regression (19) and the positive interest elasticity of the regressions of table 1 support Laidler and Parkin's main conclusion that simple long-run formulations of the demand for money do not provide a good fit to the disequilibrium quarterly data for the United Kingdom.

Next, we consider Syring's (1966) estimates of the money demand function for Norway from annual data for the period 1947–63. Syring performed the three regressions presented in table 2 as rows 2.1, 2.3, and 2.5.[4] From these regressions, Syring concluded that the demand for money was not homogeneous of degree zero in prices because R^2 drops considerably when the assumption of homogeneity of degree zero in prices is made (eq. [2.5] compared with eq. [2.3]). He also concludes that

---

[4] All data are taken from appendix A of Syring. The money supply is M_1 and consists of the annual average quantity of currency, demand deposits at commercial and savings banks, and postgiro deposits. The income variable is gross national product, the interest rate is an annual average of the long-term government bond yield, and the price index is the implicit deflator for gross national product.

TABLE 2

NORWEGIAN EQUATIONS BASED ON ANNUAL DATA FROM 1947 TO 1963
EQUATIONS $Z_{kt} = \alpha_k + \delta_k X_{kt} + \gamma_k \ln (r)_t + U_{kt}$

EQUATION No.	VARIABLES	METHOD	COEFFICIENT (STANDARD ERROR IN PARENTHESES) α_k	δ_k	γ_k	R^2	SSR	$v \approx F(3, 13)$
2.1	$Z_{kt} = \ln (M/NP)_t$	Regression	5.389 (1.491)	0.264 (0.173)	-0.239 (0.051)	.793	0.03291	...
2.2	$X_{kt} = \ln (Y/NP)_t$ $(k = 1)$	Constraint	12.440	-0.560	0.033	...	0.1001	9.540
2.3	$Z_{kt} = \ln (M/P)_t$	Regression	12.643 (3.101)	0.423 (0.131)	-0.233 (0.051)	.630	0.03187	...
2.4	$X_{kt} = \ln (Y/P)_t$ $(k = 2)$	Constraint	12.275 (3.049)	0.438 (0.132)	-0.239	...	0.03190	0.047
2.5	$Z_{kt} = \ln (M)_t$	Regression	8.508 (1.547)	0.590 (0.067)	-0.135 (0.052)	.961	0.02057	...
2.6	$X_{kt} = \ln (Y)_t$ $(k = 3)$	Constraint	5.569 (1.531)	0.718 (0.066)	-0.239	...	0.02651	1.272

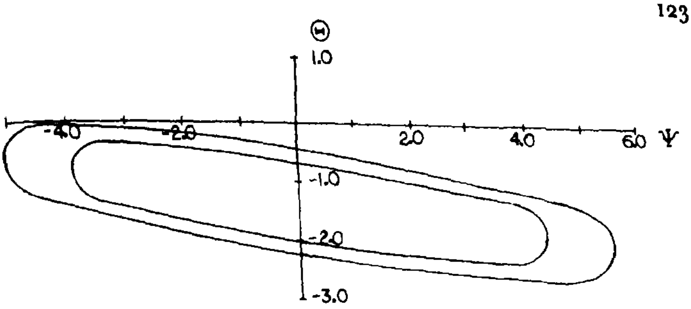


FIG. 2

the demand for money is not homogeneous of degree zero in population, because deflating by population reduces the statistical significance of most of the parameters. Finally, Syring concludes that the demand function expressed in nominal terms provides a better explanation of the data because of its higher R^2 and higher income elasticity with a lower standard error. Rows 2.4 and 2.6 of table 2 contain theoretical estimates, computed from equations (3)–(15), of the parameters and standard errors of regressions 2.3 and 2.5. The F-test presented in the last column indicates that there is no significant difference between the theoretical and regression equations. The differences between the regressions do not support Syring's conclusions and have nothing to do with the degree of homogeneity with respect to prices and population. Syring obtains a larger apparent income elasticity from nominal data because $\hat{\delta}_3$ is biased toward unity compared with $\hat{\delta}_1$. He also obtains a higher R^2 and lower standard error for $\hat{\delta}_3$ from the regression on nominal data because of the "stretching out" of the data due to the common time trends of the independent and dependent variables. The more general form of the money demand function given by equation (16) has also been estimated as

$$\begin{array}{lllllll} Z_{1t} = & -\ 6.960 & +\ 1.164X_{1t} & +\ 0.033 \ln (r)_t & +\ 0.26V_t & -\ 1.392W_t, & \\ & (17.97) & (0.39) & (0.06) & (1.39) & (0.25) & (20) \end{array}$$

$$R^2 = .946 \quad \text{and} \quad SSR = 0.0085.$$

Row 2.2 of table 2 shows the theoretical estimate of row 2.1 computed from regression (20) and equations (17) and (18). The F-test in the last column indicates that the two equations are not equivalent, and our simple model breaks down in this instance. The confidence regions for ψ and Θ are shown in figure 2. We conclude that the demand for money is not homogeneous of degree zero in prices and that the homogeneity with respect to population is indeterminate over a wide range. The range of ψ and Θ in the 95 percent confidence region gives a range on δ^* of 0.02–2.31. Again, we see that the income elasticity is poorly determined, although the apparent income elasticity from the three regressions of table 2 is relatively well determined.

Finally, we will consider several estimates of the demand for money in the United States based on yearly wealth and permanent income data from 1900 to 1958. The data interval was selected to agree with the data interval chosen by Meltzer (1963) in his work on the demand for money in the United States. For the first set of regressions considered, the variable Y is wealth, P is permanent prices, and M is currency plus demand deposits.[5] From a regression similar to equation (16), Meltzer concluded that the per capita demand for real money balances was not homogeneous of degree zero in population.[6] This implied that the demand function should be estimated in aggregate real or nominal terms, since estimating the demand function in real per capita terms would overstate the wealth elasticity. To investigate Meltzer's conclusions, we consider regressions of the three forms of the money demand function given by rows 3.1, 3.3, and 3.5 of table 3. The theoretical estimates of regressions 3.3 and 3.5, given by rows 3.4 and 3.6, indicate that there is no significant difference between the fitted and theoretical equations and that differences in the three regressions merely reflect common trends in the data. The more general form of the money demand function given by equation (16) has also been estimated as

$$\begin{array}{l} Z_{1t} = \underset{(3.20)}{3.160} + \underset{(0.16)}{1.097}X_{1t} - \underset{(0.10)}{0.812} \ln (r)_t - \underset{(0.16)}{0.296}V_t + \underset{(0.13)}{0.153}W_t, \qquad (21) \\ R^2 = .971 \quad \text{and} \quad SSR = 0.3349. \end{array}$$

Row 3.2 of table 3 gives the theoretical estimate of regression 3.1. The F-test indicates that row 3.1 can be derived from regression (21) by a simple transformation of variables. Figure 3 gives the 95 and 99 percent confidence regions for ψ and Θ. We see that Meltzer's contention that the demand function is not homogeneous of degree zero in population cannot be substantiated. The range of ψ and Θ in the 95 percent confidence region gives a range on δ^* of 0.74–1.46. Again, we see that the

---

[5] The quantity M is series X267 of *Historical Statistics of the United States: Colonial Times to 1957* (U.S. Department of Commerce 1960). Prior to 1950, wealth is defined as column 1 of Goldsmith's (1955) table W-1 minus government structures, inventories, public land, and the monetary gold and silver stock (cols. 9, 17, 18, and 24) plus the total debt of state, local, and federal governments (tables G-21, L-2, F-7, and F-19). After 1950, the wealth data are taken from Goldsmith (1962) and the state, local, and federal government debt are taken from Goldsmith and Lipsey (1964), Interest rates are 20-year corporate bond rates and are obtained from series X346 of *Historical Statistics* (U.S. Department of Commerce 1960). Permanent prices were supplied by Anna J. Schwartz.

[6] Meltzer expands equation (1) to obtain $Z_{3t} = \alpha_1 + \delta_1 X_{1t} + \gamma_1 \ln (r) + V_t + W_t + U_{1t}$. He then runs a regression on the above equation to see whether the coefficients of V_t and W_t are equal to unity. Since Meltzer is interested in aggregate formulations, a coefficient of unity implies that the aggregate demand function is homogeneous of degree 1 in the corresponding variable. This is equivalent to our real per capita formulation being homogeneous of degree zero.

TABLE 3

UNITED STATES WEALTH EQUATIONS BASED ON ANNUAL DATA FROM 1900 TO 1958
EQUATIONS $Z_{kt} = \alpha_k + \delta_k X_{kt} + \gamma_k \ln (r)_t + U_{kt}$

EQUATION No.	VARIABLE	METHOD	COEFFICIENT (STANDARD ERROR IN PARENTHESES)			R^2	SSR	$v \approx F(3, 56)$
			α_k	δ_k	γ_k			
3.1	$Z_{kt} = \ln (M/NP)_t$	Regression	−2.425 (0.421)	1.093 (0.045)	−0.741 (0.058)	.969	0.3576	...
3.2	$X_{kt} = \ln (Y/NP)_t$ $(k = 1)$	Constraint	−1.907	1.041	−0.812	...	0.3688	0.586
3.3	$Z_{kt} = \ln (M/P)_t$	Regression	−2.793 (0.722)	1.043 (0.025)	−0.759 (0.056)	.985	0.3656	...
3.4	$X_{kt} = \ln (Y/P)_t$ $(k = 2)$	Constraint	−3.039 (0.758)	1.051 (0.025)	−0.741	...	0.3666	0.048
3.5	$Z_{kt} = \ln (M)_t$	Regression	−2.299 (0.400)	1.025 (0.014)	−0.771 (0.052)	.993	0.3626	...
3.6	$X_{kt} = \ln (Y)_t$ $(k = 3)$	Constraint	−2.484 (0.453)	1.031 (0.015)	−0.741	...	0.3649	0.124

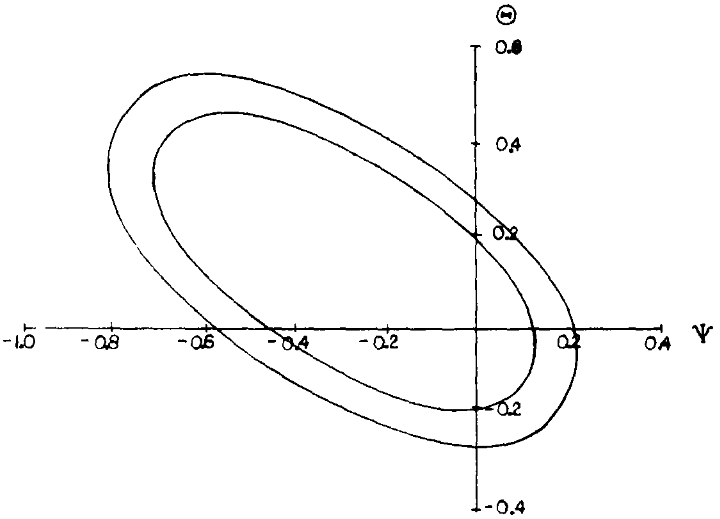


FIG. 3

wealth elasticity is poorly determined compared with the apparent wealth elasticity of the three regressions of table 3.

For the last set of regressions to be considered, the variable Y is permanent income, P is permanent prices, and M is currency and demand deposits plus time deposits.[7] Regressions of the three forms of the money demand function are given as rows 4.1, 4.3, and 4.5 of table 4. The large value for δ_1 is the type of evidence which led Friedman (1959) to conclude that real money balances are a luxury. Meltzer contends that it is Friedman's process of deflating by population which contributes to the high income elasticity and suggests that the demand function should be estimated in aggregate real or nominal terms to obtain a correct estimate of the income elasticity of money demanded. Theoretical estimates of the last two regressions are given by rows 4.4 and 4.6 of table 4. The F-test in the last column of each row indicates that Meltzer is incorrect, since differences in the three regressions have nothing to do with the degree of homogeneity with respect to population. A regression of the more general demand function is given by

$$\begin{array}{lllllll} Z_{1t} = & -7.899 & + 1.112X_{1t} & - 0.395 \ln (r)_t & + 0.382V_t & - 0.0048W_t, & \\ & (2.637) & (0.15) & (0.08) & (0.14) & (0.11) & (22) \end{array}$$

$$R^2 = 0.9776 \quad \text{and} \quad SSR = 0.2626.$$

---

[7] Permanent income and price data were supplied by Anna J. Schwartz. Interest rates are 20-year corporate bond rates and are obtained from series X346 of *Historical Statistics* (U.S. Department of Commerce 1960). Population is series A106, and the money supply is series B111-12 of *Long Term Economic Growth, 1860–1965* (U.S. Department of Commerce 1966).

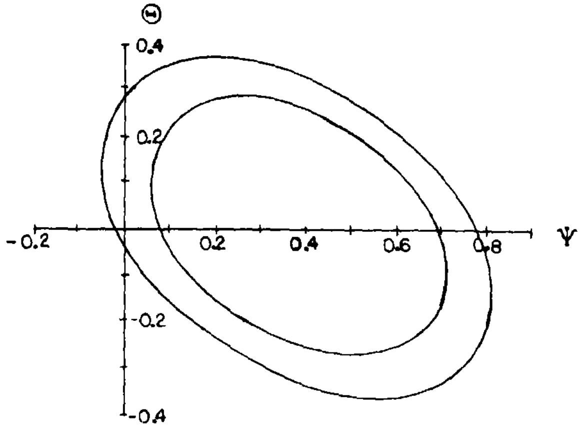


FIG. 4

Row 4.2 of table 4 gives the theoretical estimate of regression 4.1. The F-test indicates that regression 4.1 can be derived from regression (22) by a simple transformation of variables. Figure 4 gives the 95 and 99 percent confidence regions for ψ and Θ. The range of ψ and Θ in the 95 percent confidence region gives a range on δ^* of 0.77–1.45.

Conclusions

In the preceding sections we have developed a theoretical perspective of aggregation techniques employed in the estimation of the long-run demand for money from time-series data and have compared the theoretical model with data from three countries. Our theoretical model indicated that three common methods of defining the aggregate money demand function—deflating the aggregate data by population and prices, deflating by prices only, or using nominal data undeflated by population or prices—are mathematically equivalent when the data are dominated by time trend. Differences in the three aggregation procedures have nothing to do with the degree of homogeneity in prices and population, as often claimed, but merely reflect the common time trend of the variables. There appears to be no justification for estimating the long-run demand function in aggregate real or nominal terms. Although these estimates have been shown to be equivalent to the estimate based on real per capita data, they can be quite misleading if not interpreted properly, since they bias the apparent income elasticity toward unity and indicate a smaller standard error for the apparent income elasticity.

Estimates of the demand for money in real per capita terms are based on the theoretical assumption that the per capita demand for money is homogeneous of degree zero in population and prices. If this assumption is not valid, the real per capita formulation yields an estimate of the

TABLE 4

UNITED STATES PERMANENT INCOME EQUATIONS BASED ON ANNUAL DATA FROM 1900 to 1958
EQUATIONS $Z_{kt} = \alpha_k + \delta_k X_{kt} + \gamma_k \ln (r)_t + U_{kt}$

EQUATION No.	VARIABLES	METHOD	COEFFICIENT (STANDARD ERROR IN PARENTHESES)			R^2	SSR	$v \approx F(3, 56)$
			α_k	δ_k	γ_k			
4.1	$Z_{kt} = \ln (M/NP)_t$	Regression	−2.773 (0.318)	1.410 (0.042)	−0.369 (0.052)	.974	0.3067	. . .
4.2	$X_{kt} = \ln (Y/NP)_t$ $(k = 1)$	Constraint	−2.856	1.428	−0.395	. . .	0.3122	0.335
4.3	$Z_{kt} = \ln (M/P)_t$	Regression	−5.921 (0.580)	1.233 (0.022)	−0.381 (0.047)	.989	0.2694	. . .
4.4	$X_{kt} = \ln (Y/P)_t$ $(k = 2)$	Constraint	−5.683 (0.664)	1.223 (0.023)	−0.369	. . .	0.2724	0.201
4.5	$Z_{kt} = \ln (M)_t$	Regression	−3.132 (0.330)	1.126 (0.012)	−0.451 (0.045)	.995	0.2794	. . .
4.6	$X_{kt} = \ln (Y)_t$ $(k = 3)$	Constraint	−3.405 (0.369)	1.132 (0.015)	−0.369	. . .	0.2973	1.190

income elasticity which is a function of the true income elasticity and the degree of homogeneity in prices and population. Attempts to test the degree of homogeneity are not very discriminating when the data are dominated by time trend, because of the near collinearity of the variables. In the data examined, the degree of homogeneity in prices and population could vary over a considerable range, which indicates that the income elasticity was poorly determined. Since the more general form of the demand equation is poorly determined because of near collinearity, it seems justified to use the real per capita formulation if the given data set indicates that the confidence region for ψ and Θ contains the origin or if evidence from other data sources not dominated by time trend indicates that the zero-homogeneity assumption is not invalid. If a set of data should indicate that the demand for money is not homogeneous of degree zero in prices and population, we should not perform a regression on undeflated aggregate data but should take explicit account of the lack of homogeneity.

References

Central Statistical Office. *Annual Abstract of Statistics*. London: Her Majesty's Stationery Office, 1968. (*a*)

———. *Economic Trends*. London: Her Majesty's Stationery Office, October 1968. (*b*)

Chow, Gregory C. "On the Long-Run and Short-Run Demand for Money." *J.P.E.* 74, no. 2 (April 1966): 111-31.

Fisher, D. "The Demand for Money in Britain: Quarterly Results, 1951 to 1967." *Manchester School Econ. and Soc. Studies* (December 1968), pp. 329-44.

Friedman, Milton F. "The Quantity Theory of Money—a Restatement." In *Studies in the Quantity Theory of Money*, edited by Milton F. Friedman. Chicago: Univ. Chicago Press, 1956.

———. "The Demand for Money: Some Theoretical and Empirical Results." *J.P.E.* 67, no. 4 (August 1959): 327-51.

Goldsmith, R. W. *A Study of Saving in the United States*. Princeton, N.J.: Princeton Univ. Press, 1955.

———. *The National Wealth of the U.S. in the Postwar Period*. Princeton, N.J.: Princeton Univ. Press, 1962.

Goldsmith, R. W., and Lipsey, R. W. *Studies in the National Balance Sheet of the U.S.* Princeton, N.J.: Princeton Univ. Press, 1964.

Laidler, David, and Parkin, Michael. "The Demand for Money in the United Kingdom, 1956-1967: Preliminary Estimates." *Manchester School of Econ. and Soc. Studies* (June 1971), pp. 187-210.

Meltzer, Allen H. "The Demand for Money: The Evidence from the Time Series." *J.P.E.* 71, no. 3 (June 1963): 219-46.

Syring, Edward M. *The Demand for Money in Norway*. Oslo: Universitetsforlagets Trykmingssentral, 1966.

U.S. Department of Commerce. *Historical Statistics of the United States: Colonial Times to 1957*. Washington: Bur. Census, 1960.

———. *Long Term Economic Growth, 1860-1965*. Washington: Bur. Census, 1966.

Communications

External Economies and the Production Frontier

J. R. Gould

London School of Economics

This paper sets out an analysis of a simple, two-good general equilibrium model in which one industry confers an external benefit on the other. It shows that, except in special cases, the locus of competitive equilibrium combinations of goods has both negatively and positively sloping branches. Several interesting results stem from this point. Unstable equilibria are possible which arise on the supply side, so to speak, rather than, as in the standard simple general equilibrium model, from the opposition of income and substitution effects in consumption. Second, in standard models, equilibrium occurs on the production frontier unless marginal rates of factor substitution are not equalized between industries; in this single-factor model, equilibrium can be off the production frontier, notwithstanding that inefficient factor combinations are ruled out by assumption. Third, shifts in preferences from one good to the other can decrease the equilibrium outputs of *both* goods. Finally, and somewhat paradoxically, a unique and stable equilibrium is possible with zero outputs of both goods.

The model is readily adapted to the analysis of an external diseconomy, and it would be tedious to repeat the analysis. For purposes of contrast, the main results for the external diseconomy case are briefly summarized in the concluding section.

The Model

The model is a modification of the timber-wheat case discussed by Meade (1952). A composite factor l, which will be called "land," is in fixed supply and can be used in the production of x_1 (wheat) or x_2 (timber). The resource constraint is $l_1 + l_2 = l$. The output of timber depends on the amount of land applied to timber growing; wheat output depends both on the amount of land devoted to wheat and on the

---

I would like to thank C. Blackorby, B. V. Hindley, H. G. Johnson, and M. Perlman for helpful comments on a previous draft of this paper.

[*Journal of Political Economy*, 1974, vol. 82, no. 6]

© 1974 by The University of Chicago. All rights reserved.

amount of rainfall; rainfall increases with the amount of timber grown. The production functions can be written $x_1 = x_1(l_1, x_2)$ and $x_2 = x_2(l_2)$. The assumptions about the marginal products of the x_1 production function are: the marginal product of l_1 is constant with respect to l_1 and increases with x_2; the marginal product of x_2 diminishes as x_2 increases and increases with l_1. The x_2 production function is subject to constant returns. All marginal products are assumed to be positive.

In order to simplify the demand side as much as possible, it is assumed that all consumers own an equal share of the land, that all have identical preferences, and that neither x_1 nor x_2 is an inferior good.

Wheat, timber, and land are sold in perfectly competitive markets. Rainfall is an externally supplied good, since timber growers cannot charge wheat growers for rain. Although this may seem to be stating the obvious, assumptions about the existence and feasibility of enforcement of property rights are best not left implicit in discussions of external economies. If a scarce good is unpriced, it is fairly obvious that it will be supplied inefficiently; what is often overlooked is that it may, or may not, be *used* inefficiently.[1] Thus, if the rain were to seep through to a water table from which all wheat farmers pumped water to irrigate their crops, it would be a common-property resource and it would be exploited inefficiently. If, on the other hand, the rain were captured by the topsoil, it would be the private property of the wheat farmers on whose land it fell and it would be exploited efficiently. In order to focus attention on the supply effects of the external economy, it will be assumed that rain is captured by the land on which it falls and is used efficiently.

The Production Frontier

Figure 1 is a diagrammatic derivation of the production frontier from the two production functions and the land constraint. In figure 1*a*, the isoquants x_1' and x_1'' represent the x_1 production function. The straight line GG' shows the relation between x_2, the output of timber and l_1, the quantity of land employed in the wheat industry, given that the total supply of land l is fully employed. Alternatively, taking the point l_1 as origin and measuring l_2 westward, GG' simply relates total product of x_2 to l_2.

The curve $FF'F''$ in figure 1*b* is derived from the intersections of GG' with the x_1 isoquants. Thus, if l_1' units of land are employed in wheat and $l - l_1'$ in timber, timber output is x_2''; the combination of l_1' and x_2'' produces x_1' units of wheat. The output of wheat is a maximum of x_1'' when l_1'' units of land are employed in the wheat industry and the x_1

---

[1] This point was overlooked by Meade (1952), and, as a consequence, part of his analysis is erroneous (see Gould 1973).

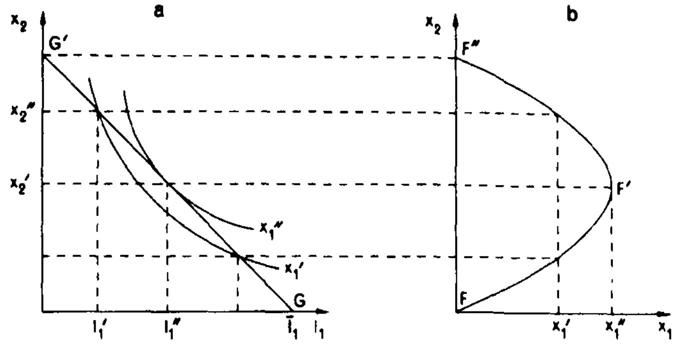


Fig. 1

isoquant is tangent to GG'. As the input of land into wheat is increased beyond l_1'', the outputs of both timber and wheat decline.

Defined as the locus of output combinations such that it is impossible to produce more of one good without producing less of the other, the production frontier is represented in figure 1*b* by $F'F''$. To distinguish the production frontier so defined from the curve which encloses the set of feasible output combinations, we can call the latter, $FF'F''$ in figure 1*b*, the "production boundary."

The necessary and sufficient condition for the outputs to be on the production boundary $FF'F''$ is simply that land be fully employed. The condition for an intersection to generate a point on the production frontier $F'F''$ is that GG' cut the x_1 isoquant from below; that is, the slope of GG' must be less in absolute value than the slope of the isoquant:

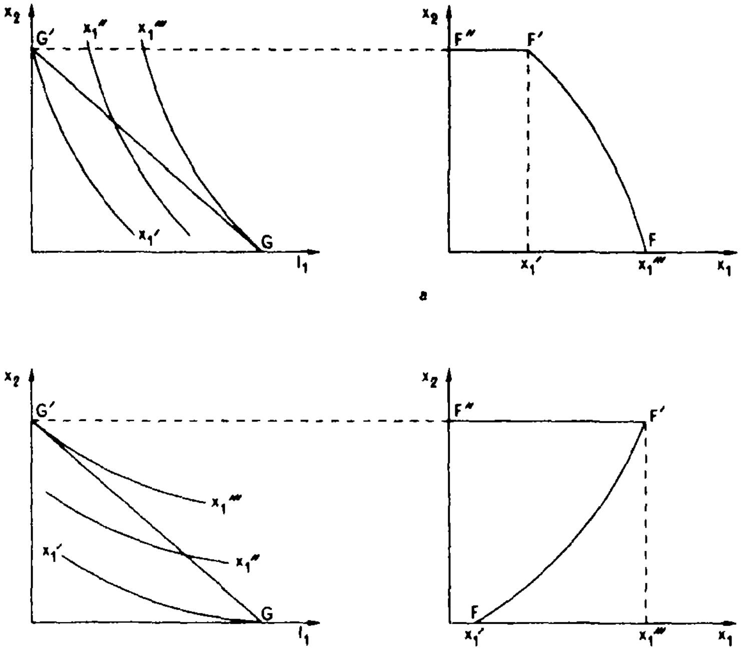
$$\frac{dx_2}{dl_2} < \frac{\partial x_1/\partial l_1}{\partial x_1/\partial x_2}.$$

This can be written in a form which has a commonsense interpretation:

$$(\partial x_1/\partial x_2)(dx_2/dl_2) < \partial x_1/\partial l_1.$$

The right-hand side of the inequality is the direct marginal contribution of land to wheat production; the left-hand side is the indirect marginal contribution via timber-induced rainfall. Plainly, if the indirect exceeds the direct contribution, it is possible to shift land from wheat to timber and increase the outputs of both goods.

If both l_1 and x_2 are essential to the production of x_1, the x_1 isoquants do not touch the axes, there must be an interior tangency at which output of x_1 is a maximum, the production boundary must have an upward-sloping branch and a downward-sloping branch, and F coincides with



a

b

Fig. 2

the origin. If either l_1 or x_2 are not essential factors, an interior tangency is still possible but no longer necessary. Two corner solutions are illustrated in figure 2. In figure 2*b* the production frontier reduces to the single point F'.

The curvature of the production boundary can be deduced, in this simple model, by inspection of the marginal rate of transformation, *MRT*, expressed in terms of marginal products. Taking total differentials of the production functions and the resource constraint and manipulating,

$$MRT = \frac{dx_2}{dx_1} = \frac{dx_2/dl_2}{[(\partial x_1/\partial x_2)(dx_2/dl_2)] - (\partial x_1/\partial l_1)}.$$

In passing, note that $MRT < 0$ is a condition for outputs to be on the production frontier, from which the inequality condition previously derived can be confirmed. Now consider a move down the production boundary, which necessarily entails decreases in l_2 and x_2, and an increase in l_1. From the specifications of the production functions, dx_2/dl_2 remains constant; and as l_1 increases and x_2 decreases, $\partial x_1/\partial l_1$ decreases and $\partial x_1/\partial x_2$ increases. Thus, the numerator of the *MRT* stays constant and the denominator increases. For movements down the

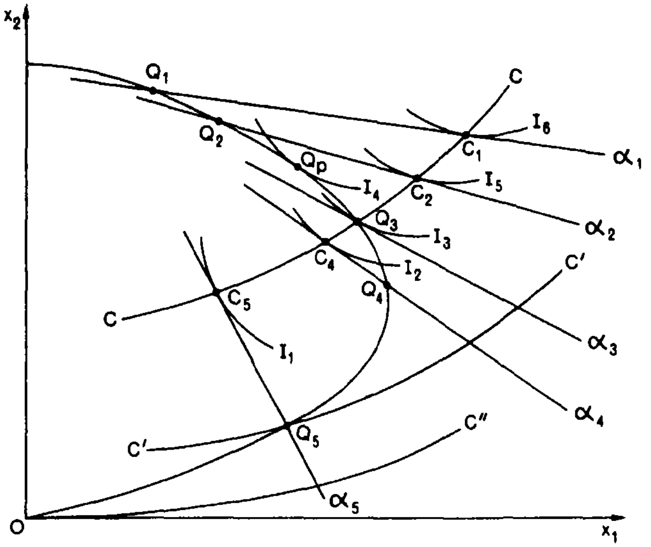


FIG. 3

production boundary, the algebraic value of MRT decreases[2] and, geometrically, the production boundary is concave to the x_2 axis.

Competitive Equilibrium

Given that the marginal products of land are always positive, equilibrium in the market for land implies that land is fully employed. Thus, the locus of equilibrium combinations of x_1 and x_2 must lie on the production boundary $FF'F''$.

The private marginal costs of x_1 and x_2 are $R/(\partial x_1/\partial l_1)$ and $R/(dx_2/dl_2)$, where R is the price of land. The ratio of the private marginal costs $(dx_2/dl_2)/(\partial x_1/\partial l_1)$ differs from the social marginal cost ratio—the absolute value of MRT—because timber's indirect contribution to wheat production is external to the price mechanism: the private marginal cost of timber in terms of wheat exceeds its social marginal cost.

In figure 3, the private marginal cost ratio at a given point on the production boundary is indicated by the absolute value of the slope of the α line through that point. As the production point moves down the production boundary, the slopes of the α lines become increasingly steep because dx_2/dl_2 remains constant and $\partial x_1/\partial l_1$ decreases as x_2

---

[2] With reference to fig. 1*b*, for movements from F'' to F', MRT is negative and diminishing, tending to minus infinity at F'; at F' there is a discontinuity; and for movements from F' to F, MRT is positive and diminishing from plus infinity.

decreases and l_1 increases; and, although the social marginal cost of timber in terms of wheat can be negative, its private marginal cost must be positive.

For firms to be in equilibrium, the goods price ratio P_1/P_2 must equal the private marginal cost ratio. Thus, the supply side of the economy is in equilibrium if the combination of x_1 and x_2 is on the production boundary and the price ratio equals the slope of the relevant α line.

On the demand side, the assumption that all consumers have equal endowments of land and the same preferences permits the use of an aggregate preference function and an aggregate budget constraint. In figure 3, the aggregate preference function is represented by the indifference curves I_1, I_2, etc. The position and slope of the budget constraint depend on the outputs of the economy. If, for example, the supply side of the economy is in equilibrium at Q_1, the aggregate budget constraint passes through Q_1 and its slope reflects the price ratio, which in turn is equal to the ratio of the private marginal costs. Thus, the line α_1 produced to the axes can be interpreted as the aggregate budget constraint facing consumers if the supply side of the economy is in equilibrium at Q_1.

Given the budget constraint α_1, consumers will choose to purchase the combination of x_1 and x_2 at C_1. Similarly, if the production point is Q_2, the budget constraint is α_2 and the chosen consumption point is C_2. The line CC is the locus of such equilibrium consumption points.

The direction in which the associated consumption point moves as the production point moves down the production boundary depends on the balance between income and substitution effects. Since the relative price of x_1 rises, there are substitution effects against x_1 and in favor of x_2. The income effects are a little less straightforward.

If two successive consumption points are to the southeast of their respective production points (e.g., C_1 and C_2), the one corresponding to the production point lower on the production boundary must be on a lower indifference curve, since the α lines diverge toward the southeast. Given that neither x_1 nor x_2 are inferior goods, the income effects reinforce the substitution effect against x_1 and counteract the substitution effect in favor of x_2: consumption of x_1 must fall, and that of x_2 may rise or fall. Thus, the consumption point must travel toward the northwest or southwest as the production point moves down the production boundary.

On the other hand, if two successive consumption points are to th northwest of their corresponding production points (e.g., C_4 and C_5) since the α lines converge toward the northwest and may intersect in th positive quadrant, the second consumption point may be on a higher c lower indifference curve than the first. If the move is to a lower indiffe ence curve, then, again, the direction of the consumption point toward the northwest or the southwest. If the move is to a higher indi ference curve, the substitution effect against x_1 is counteracted and th in favor of x_2 is reinforced: the consumption point travels either toward the northwest or the northeast as the production point moves down the production boundary. The one direction in which the consumption point cannot travel is toward the southeast.

In figure 3 there is a unique and stable general equilibrium at Q_3, with demand equal to supply for both x_1 and x_2. Suppose the equilibrium price ratio indicated by α_3 is disturbed, P_1 falling relative to P_2. The production point moves up the production boundary, generating excess demand for x_1 and excess supply of x_2, as illustrated by the production point Q_2 and the corresponding consumption point C_2. Conversely, a rise in P_1 relative to P_2 would generate excess supply of x_1 and excess demand for x_2, illustrated by Q_4 and C_4.

There are no market forces which rule out the possibility of equilibrium on the upward-sloping branch of the production boundary. The point is perhaps best illustrated by considering a shift of preferences from x_2 to x_1, given an initial equilibrium on the production frontier. In figure 3, the new preferences (not represented by indifference curves in the diagram) are such that the new locus of consumption points is $C'C'$. The shift in preference from timber to wheat creates excess supply of timber and excess demand for wheat at the initial equilibrium production point Q_3, and the price of wheat rises relative to the price of timber. Producers attempt to increase the output of wheat and to decrease the output of timber, and, in the process, wheat farmers bid land away from timber producers. However, after a point, further reductions in timber output so diminish rainfall that the output of wheat declines, notwithstanding that more land is devoted to its production. Although all wheat farmers may realize what is happening, no individual farmer in atomistic competition has the incentive to do other than attempt to increase his output of wheat by hiring more land as long as the price of wheat exceeds his private marginal cost. Eventually, an equilibrium consistent with the new preferences is reached at Q_5. If we compare initial and final equilibrium points in this example, the shift in preferences from timber to wheat has reduced the output of both.

As an extreme case, there is the curious possibility that the economy can have a unique and stable equilibrium with zero outputs of both goods. Suppose preferences are such that the locus of consumption points lies outside the production boundary except at the origin, as illustrated by OC'' in figure 3. The origin is a position of equilibrium because demands and supplies of x_1 and x_2 are all equal to zero; and at any of the other possible production points, there is excess demand for x_1 and excess supply of x_2, so that the equilibrium is unique and stable. A necessary condition for this result is that x_2 is an essential factor in the production of x_1: the intercept of the production boundary with the x_1 axis must be at the origin rather than at some positive quantity of x_1.

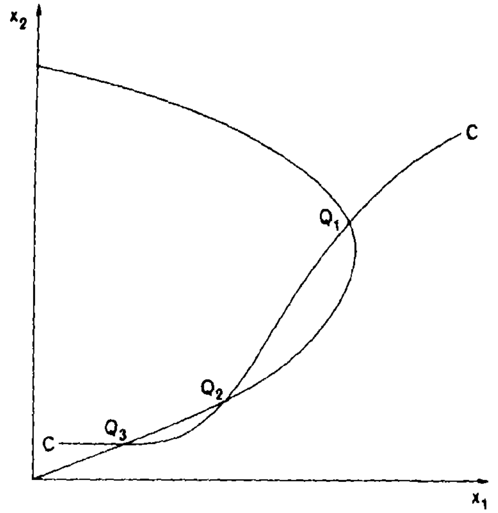


FIG. 4

Given that neither x_1 nor x_2 are inferior goods, there can be at most one equilibrium on the production frontier. This is because, once the locus of consumption points has crossed the production frontier from the right, it would, at some stage, have to travel toward the southeast to generate another equilibrium intersection with the production frontier, but this is the one direction which is ruled out in the discussion of the locus of consumption points above. However, the upward-sloping branch of the production boundary gives rise to the possibility of multiple and unstable equilibria. Figure 4 illustrates this, with Q_1 and Q_3 being positions of stable equilibrium and Q_2 one of unstable equilibrium.

Economic Welfare

An equilibrium on the upward-sloping branch of the production boundary is obviously inefficient. A small subsidy on timber would increase the outputs of both timber and wheat, and, in this simple model, all consumers would be better off.

Even if it occurs on the production frontier, equilibrium is not Pareto optimal. The marginal rate of substitution in consumption differs from the marginal rate of transformation in production because market forces bring the price ratio into equality with the private, rather than the social, marginal cost ratio. As illustrated in figure 3, relative to the Pareto-optimal production point, Q_p, too much wheat and too little timber are produced at the equilibrium production point Q_3.

A Pareto optimum can be induced by a Pigouvian subsidy to timber producers.[3] Let s_2 be the required ad valorem subsidy on x_2. Profit-maximizing firms equate price plus subsidy to private marginal cost. Thus,

$$\frac{P_1}{P_2(1 + s_2)} = \frac{dx_2/dl_2}{\partial x_1/\partial l_1}. \tag{1}$$

For a Pareto optimum, the slope of the budget constraint facing consumers must equal the marginal rate of transformation:

$$\frac{-P_1}{P_2} = \frac{dx_2/dl_2}{[(\partial x_1/\partial x_2)(dx_2/dl_2)] - (\partial x_1/\partial l_1)}. \tag{2}$$

Combining (1) and (2),

$$s_2 = \frac{(\partial x_1/\partial x_2)(dx_2/dl_2)}{(\partial x_1/\partial l_1) - [(\partial x_1/\partial x_2)(dx_2/dl_2)]},$$

to be evaluated at the Pareto-optimal production point, Q_p.

External Diseconomies

It is apparent that the model can be adapted to the analysis of external diseconomies simply by changing the sign of $\partial x_1/\partial x_2$. In the revision of figure 1, the x_1 isoquants would slope upward and the production boundary and the production frontier would coincide and slope downward. Inspection of the expression for the marginal rate of transformation shows that, moving down the production boundary, the sign of the change in *MRT* can be either positive or negative: the production boundary may be either convex or concave to the origin.[4] Equilibrium, not necessarily stable or unique, is on the production frontier, but because the private marginal cost of x_2 in terms of x_1 is greater than its social marginal cost, too much x_2 is produced. A Pigouvian tax on x_2 can induce a Pareto optimum.

References

Baumol, W. J., and Bradford, David F. "Detrimental Externalities and Non-Convexity of the Production Set." *Economica* 39 (May 1972): 160–76.

Gould, J. R. "Meade on External Economies: Should the Beneficiaries be Taxed?" *J. Law and Econ.* 16 (April 1973): 53–66.

Meade, J. E. "External Economies and Diseconomies in a Competitive Situation." *Econ. J.* 62 (March 1952): 54–67.

---

[3] Since there is only one price ratio to be adjusted in this model, a subsidy on land used in timber production or a tax on wheat or land used in wheat production could achieve the same end. However, with more goods and factors, taxes on wheat or on land used in wheat production would distort price ratios between wheat and other goods. The choice would be restricted to a subsidy on timber or an ad valorem subsidy at the same rate on all factors employed in timber production.

[4] See also Baumol and Bradford (1972), who discuss a similar model.

The Staple Theory and Its Empirical Measurement

Dwight Grant

University of Alberta

I. Introduction

Our interest in this subject centers upon an earlier analysis by Chambers and Gordon.[1] Their objective was to develop a model which would permit measurement of the effects of staple production upon economic growth. Begun by Harold Innis and continued by others,[2] the staple theory argues that economic growth can, in some countries, be attributed to the leading influence of staple exports. Canada is a prominent historical example of such a country. Of Canada's staples, wheat in the prairie west during the boom period 1900–1915 is considered the outstanding example. Chambers and Gordon sought to measure the influence of prairie agricultural production on Canadian per capita income.[3] Their analysis was innovative and fruitful; their measurements indicated that a low proportion of the per capita income increase experienced in this "golden age in the economic growth of Canada" was attributable to prairie wheat.[4] They calculated that between 6.10 and 9.30 percent of intensive growth was so attributable.[5] In this paper we accept the fundamental elements of their argument, but we contend that they have overlooked an important factor, entrepreneurial risk bearing, which affects their estimates and conclusions materially. We will present our contention and then compute a revised set of estimates of the measure of the influence of wheat on per capita Canadian income for this boom period.

At an early stage in this research I benefited from the comments of Irving Kravis, and, later, from those of an anonymous referee.

[1] See Chambers and Gordon 1966.

[2] See Watkins 1963.

[3] We leave to others the debate on whether extensive rather than intensive growth was the subject of the historian's staple theory. Intensive growth, and measurement of the extent to which factors contribute to it, is a worthwhile enterprise, in its own right.

[4] Chambers and Gordon 1966, p. 316.

[5] These are the high and low estimates which appeared in Chambers and Gordon (1967). The difference is attributable to alternative assumptions about induced population changes.

[*Journal of Political Economy*, 1974, vol. 82, no. 6]

© 1974 by The University of Chicago. All rights reserved.

II. A Change in the Model

Land, labor, and, at a second stage, capital, were the factors of production in the Chambers and Gordon model. The price of labor for the economy was established by the value of its marginal product in a domestic gadget industry. The wheat industry therefore determined its output by employing labor just until its marginal value product no longer exceeded its cost. The difference between the total value product of wheat and its total labor input cost represented land's rent—the value produced in excess of the economic opportunity cost of the labor input. Capital, or any other factor input, could be added to this model without an adverse impact, "provided these factors are available to industries (from Canada or from abroad) at fixed world prices, or at prices fixed by the world market plus transportation and/or tariff."[6] In such a situation, and Chambers and Gordon thought this a reasonable representation of the Canadian economy circa 1910, the contribution of the prairie wheat boom to intensive Canadian growth could be measured as the total of the increase in land rent value above what would otherwise have obtained and the increase in tariff revenue occasioned by the wheat boom. The latter increase was collected on the substitution of imports for domestic goods in the gadget industry. The aggregate influence was adjusted for the increase in population induced by the staple-led growth.

Our concern for the efficacy of the model first arose when it was calculated that, of the total 1911 prairie grain production of $223 million,[7] labor and imputed rent accounted for only $148 million,[8] or 66 percent of total value product. The estimates of Chambers and Gordon would remain valid only if the unaccounted 34 percent of value flowed to factor inputs at their opportunity cost in the gadget industry. We argue that in all likelihood it did not!

Survey data for the production season 1911 details the costs and revenues experienced by the prairie grain economy. It indicates that, in 1911, farmers in Alberta, Saskatchewan, and Manitoba enjoyed profits, after deduction for imputed rent and personal labor, of $49.9 million.[9] It is our contention that this profit figure, which accounts for an additional 22 percent of total value product, is the major omitted factor return, and that it is, furthermore, a return in excess of the economic opportunity cost of the factor inputs. Specifically it can be characterized as a return to entrepreneurial expertise and risk bearing.

---

[6] Chambers and Gordon 1966, p. 321.

[7] *Canada Yearbook, 1914*, pp. 154–63.

[8] This is based on the estimate of Chambers and Gordon (1966, p. 320), that labor costs were equal to 46 percent of total value product and that rent in 1911 was $45.8 million.

[9] Calculations are based on acreage profit figures contained in *Quarterly Bulletin of Agricultural Statistics* (1912, 1914); and on planted acreage figures from *Canada Yearbook 1914*, pp. 154–63.

In the context of the Chambers and Gordon model, we must argue not only that there was a return to entrepreneurial risk bearing, but also that equivalent opportunities did not exist elsewhere in the economy. Any claim concerning the latter point is somewhat tenuous. Nevertheless, the unique aspects of agricultural homesteading as an entrepreneurial activity and the volume of people involved (148,000 families)[10] lead us to believe that returns in the form of profits represented returns above the economic opportunity cost of the resources.

We will present no empirical substantiation of the claim that there was significant entrepreneurial expertise required in the wheat industry. The argument for the existence of risk in this industry is persuasive. We cite one example which is indicative: In 1911, the average profit from an acre of barley in Manitoba was $5.98; in 1913, the same crop yielded an average profit of $1.02.[11] The variations experienced by the individual farm unit must almost certainly have been far more extreme. In a risk-averse society, the individual owner or tenant would require a positive expected profit as compensation for risk bearing. By excluding expected profits from their estimation of the contribution of prairie agriculture to intensive growth, Chambers and Gordon omitted an important theoretical consideration and understated, substantially, agricultural influence on growth.

III. A Revised Estimate

We have no quarrel with the approach to estimation which Chambers and Gordon employed. Therefore, to the extent consistent with the addition of an expected profits figure, we will follow their estimation procedure. From the production-cost survey data, we estimated the value of rent and profit in prairie grain production in 1911. Since comparable data for 1901 do not exist, we assumed that the relationships were proportionate to those estimated for rent by Chambers and Gordon. Taking their earlier estimates of the influence of the wheat boom on population changes and tariff revenues, we combined these measurements to develop a range of estimates of the influence of wheat on intensive growth.

The imputed rent figure included in the survey data is actually a measure of interest carrying cost, that is, the market value of property multiplied by the interest rate appropriate for real estate secured loans.

---

[10] Chambers and Gordon 1966, p. 323.

[11] *Quarterly Bulletin of Agricultural Statistics* (1914), p. 305.

Whether or not this is an accurate reflection of rental value is a trifling matter, since we are interested in the aggregate of rent and profit—the residual of revenues after the deduction of all factor input costs[12] save land. The 1911 figure for rent for the three prairie provinces is $42.0 million.[13] This compares closely with the $45.8 million estimate of the earlier authors.

The aggregate profit earned in 1911 was $49.9 million; in 1913 it was $27.6 million.[14] In addition to confirming the claim that risk was an important characteristic of the agricultural enterprise, this disparity leaves us in a quandary as to what constitutes an accurate measure of profit's contribution to growth. In the absence of reasonable alternatives, we will use $49.9 million as a high figure, $27.6 million as a low figure, and their average $38.8 million as a best estimate. To estimate 1901 profit figures, we applied the high, low, and average profit margins to 1901 grain revenue.[15] Tariff revenue was estimated to have changed as the earlier authors indicated.

We assumed that the 1901 rent bore the same relationship to 1911 rent as the relationship established by the earlier authors between 1901 and 1911 rent. On this basis, the aggregate impact on income of the change in rent,[16] profits, and tariffs is in the range of $98.3 million, $87.6 million, and $76.8 million. Combining this with the alternative immigration assumptions of either no induced immigration or an increase in immigration of 200,000 people,[17] we can derive six alternative estimates of the proportion of intensive growth attributable to staple growth, as shown in table 1.

TABLE 1

ESTIMATES OF THE PROPORTION OF GROWTH

	PROFITS		
IMMIGRATION	High	Average	Low
0	0.25	0.22	0.19
200,000	0.21	0.18	0.16

---

[12] Instructions accompanying the survey made explicit the need to charge all labor at the applicable local rate, whether it was hired or performed by the owner (see *Quarterly Bulletin of Agricultural Statistics* [1912], p. 51 and [1914], p. 299).

[13] This is based on the data mentioned in n. 9 above. The corresponding figure for 1913 was $44.2 million. This small difference was ignored in later calculations.

[14] See n. 9 above. This observed volatility of profits makes it difficult to assign precise meaning to any estimate of the contribution of the industry to growth over a 10-year period.

[15] I.e., $7.22 million (see *Fourth Census of Canada 1901*, 2: 232).

[16] Our change in rent figure is $37.05 million, which was calculated as follows: 37.05 = (40.4/45.8)(42.0). The change in tariff revenue was estimated to be $13 million (Chambers and Gordon 1967, p. 882).

[17] Chambers and Gordon 1966, p. 323.

IV. Conclusion

Chambers and Gordon asserted that even a generous interpretation of the influence of prairie wheat on intensive Canadian growth leads to the conclusion that it was only marginally important. The inclusion of a return to risk-bearing entrepreneurs alters this conclusion. Our revised (and admittedly generous) range of estimates ascribes a far higher proportion of intensive growth to prairie agriculture. Even our low estimate exceeds their high estimate by a factor of almost two. On the other hand, we dispute the contention of previous critics[18] that the Chambers and Gordon model is strongly biased against measuring the influence of wheat on intensive growth. The model can and does prove useful in measuring a substantial influence when the return to entrepreneurial risk bearing is included.

References

Canada, Bureau of Statistics. "Cost of Grain Production in Canada." *Q. Bull. Agricultural Statis.* (March 1912), pp. 51–57.

———. "Cost of Grain Production in Canada." *Q. Bull. Agricultural Statis.* (December 1914), pp. 299–306.

Canada, Census Office. *Fourth Census of Canada, 1901.* 4 vols. Ottawa: King's Printer, 1902–06.

Canada, Dominion of. *Canada Yearbook, 1914.* Ottawa: King's Printer, 1915.

Chambers, E. J., and Gordon, D. F. "Primary Products and Economic Growth: An Empirical Measurement." *J.P.E.* 74, no. 4 (August 1966): 315–32.

———. "Primary Products and Economic Growth: Rejoinder." *J.P.E.* 75, no. 6 (December 1967): 881–85.

Dales, J. H.; McManus, J. C.; and Watkins, M. H. "Primary Products and Economic Growth: A Comment." *J.P.E.* 75, no. 6 (December 1967): 876–80.

Watkins, M. H. "A Staple Theory of Economic Growth." *Canadian J. Econ. Polit. Sci.* (May 1963), pp. 141–58.

---

[18] Dales, McManus, and Watkins 1967.

Risk, Job Search, and Income Distribution

Christopher A. Pissarides

London School of Economics

I. Introduction

One of the by-products of the literature which followed the publication of Phillips's (1958) well-known paper was the development and elaboration of models of individual behavior in labor-market disequilibrium (see Phelps [1969] for a survey). In retrospect this may be the most significant consequence of the Phillips curve analysis, since the labor-market models may suggest ways for tackling the wider problem of the adjustment to equilibrium which is still in a very messy state. In this paper we shall consider one aspect of the problem which has been neglected by the "new microeconomics," namely, the implications of attitudes toward risk for individual behavior in disequilibrium. Section II introduces the concept of a risk premium which is then used in Section III to analyze the relationship between individual behavior toward risk, job search, and the structure of wage offers by firms. Sections IV and V consider the relationship between the distribution of risk attitudes and the distribution of income and wealth among individuals. A concluding section points out some of the implications of the analysis for empirical research.

II. Behavior toward Risk and the Risk Premium

It has recently been shown by Pratt (1964) that the work of von Neumann and Morgenstern (1947) may be used to provide the rationale for the existence of a *risk premium*, which may be defined as that amount of money which induces an individual to give up a certain income in preference for risky returns with the same expected value as the certain income. Let W denote total wealth, and consider an individual who has the option either of a certain wealth W^0 or of a randomly distributed wealth W. The risk premium p is the difference $EW - W^0$, where EW is the expected value of W, which makes the individual indifferent between the two alternatives. Pratt suggests using the risk premium as a

I wish to thank Professor Michio Morishima for his comments and guidance during the preparation of this paper. A referee has also made useful comments for improvement.
[*Journal of Political Economy*, 1974, vol. 82, no. 6]
© 1974 by The University of Chicago. All rights reserved.

measure of local or global risk aversion and shows that it varies inversely with the willingness to bear risks.

Let $U(W)$ be a von Neumann–Morgenstern utility function defined over total wealth. If this function is (strictly) concave the individual is a risk averter, if it is linear he is indifferent to risk, and if it is convex he is a risk lover. If individuals maximize expected utility the definition of the risk premium implies

$$EU(W) = U(EW - p). \tag{1}$$

Thus by the definition of convex and concave functions the risk premium is positive for a risk averter, zero for an individual indifferent to risk, and negative for a risk lover. Pratt further shows that the risk premium varies directly with the *absolute risk-aversion coefficient* which is defined by $R = -U''/U'$ (where primes denote differentiation) and which was suggested independently by Arrow (1965) and Pratt as a measure of local risk aversion. Moreover, a straightforward extension of the Rothschild-Stiglitz (1970) definition of the degree of uncertainty of a probability distribution will show that the risk premium varies directly with the degree of uncertainty. Thus let p_1 be the risk premium associated with W_1 and p_2 the risk premium associated with W_2, where $EW_1 = EW_2$ and the utility function is unchanging. Then Rothschild and Stiglitz show that an appealing and useful way of measuring increasing riskiness is equivalent to writing

$$EU(W_1) \geq EU(W_2) \tag{2}$$

for all concave U when W_2 is more risky than W_1 and with reverse inequality sign for all convex U. Using (1) we obtain

$$U(EW_1 - p_1) \geq U(EW_2 - p_2). \tag{3}$$

By concavity, and since $EW_1 = EW_2$,

$$p_2 \geq p_1. \tag{4}$$

Thus Pratt's risk premium has all the desirable properties of the risk premium introduced in economic theory in various contexts (e.g., Keynes [1930, pp. 142–44]; Hicks [1946, pp. 137–39]), and it is consistent with the von Neumann-Morgenstern utility axioms.

III. Job Search and the Structure of Wage Offers

The common assumption of the "new microeconomics" of the labc market which distinguish them from more traditional approaches is th the labor-market participant has only imperfect information about jo offers.[1] In particular, it is usually assumed that there is a fixed distribution of wage offers in the market, and the worker knows this distribution but has no information about individual vacancies. Information about individual vacancies can only be obtained directly through search, so the worker, in seeking employment, contacts prospective employers and enquires about vacancies. If a vacancy is available and the wage rate is "acceptable" to the worker according to some criteria, then he accepts the job; if not he rejects it and continues the search. The criterion that determine the acceptable wage rate is usually maximization of expected income, given a fixed and known distribution of wage rates. The main conclusion of the analysis is that there will be some "equilibrium" unemployment which depends on the degree of labor-market imperfections and consists of people in search of employment. For this reason this equilibrium unemployment may be interpreted as frictional unemployment.

In general an individual decides to choose an employment on the basis of his evaluation of all the principal features of the employment besides pecuniary awards. Even if we ignore nonpecuniary awards, the problems associated with the expected length of employment, security of employment (the risk of being made unemployed), promotion prospects, and more generally the factors related to the uncertainty of future income, cannot be ignored in a model which puts emphasis on imperfect information and risk attitudes. In principle these factors may be taken into account in a straightforward manner by assuming that the information conveyed when a job offer is considered enables the receiver of the offer to estimate the probability of each alternative and so obtain the expected utility of the job offer. This, however, complicates the analysis enormously since it implies that a job offer cannot be described by simple objective factors, such as the current wage rate it offers or the discounted value of the income stream it promises. These complications have been ignored by the job-search literature where a single wage rate is taken to describe objectively and fully a given job offer, and an extension of the analysis along these lines is clearly desirable. The main argument of this paper however, may be brought out in a simplified model where a job offer is fully described by a single wealth variable. This wealth variable may be interpreted as the present value of the income stream a job offer promises to pay over time, adjusted for the uncertainties mentioned above through, for example, the discount rate. We shall refer to this wealth variable as "income," or "wages," to distinguish it from nonhuman wealth.

---

[1] Most of the authors of job-search models, e.g., the contributors to Phelps et al. (1970), claim to be formalizing and extending Stigler's (1962) work on labor-market imperfections. However, although Stigler is clearly responsible for the current interest in these problems, an older and in many ways clearer statement of the theoretical issues involved appears in Hutt (1939, chaps. 3–5). Hutt also considers risk attitudes more explicitly than Stigler. This makes Phelps's (1969) adjective "new" rather questionable.

Occasional remarks about the likely effects of risk on behavior abound in the job-search literature, but in the formal models it is usually assumed that the individual attempts to maximize the present value of his income stream, that is, that he is indifferent to risk. An obvious defense for using income maximization and still taking account of risk would run along Hicksian lines: Expected incomes are adjusted according to risk premia so during the search process the household attempts to maximize the *certainty-equivalent* of discounted lifetime income. In order to examine the implications of this way of dealing with risk, we make the analysis more explicit by assuming that the typical participant's criteria of acceptability are determined by expected utility maximization and that the information he has about the labor market may be summarized in a probability distribution of expected income earnings $P(Y)$, where $P(Y_i)$ is the probability of earning an income stream whose discounted present value is Y_i. Probabilities may be objective or subjective, and they may be revised when further information is collected through job search. Since those in employment have much clearer notions about their earning prospects than those unemployed, we may simplify the analysis by assuming that once an individual accepts a job offer he expects to be earning a certain income. More formally, the set of all income streams which carry a positive probability consists of a finite number of elements for those in search of employment and of a single element for those already in employment. The employment contract in such a model may be thought of as a long-term (lifetime) contract.

Consider an individual who has property wealth A and who is searching for employment, expecting to be offered a job paying income Y with probability $P(Y)$. It is assumed that Y is the only random element of total wealth W, which is equal to $A + Y$. If this individual were offered a job paying income Y_i^0 he would accept the offer if and only if

$$U(A + Y_i^0) \geq EU(A + Y). \tag{5}$$

Comparing (5) with (1) we see that this condition is equivalent to

$$Y_i^0 \geq EY - p. \tag{6}$$

So a job offer will be accepted if the mean expected wage offer does not exceed the offer by more than the risk premium. The minimum value of Y_i^0 that satisfies (6) is the minimum wage offer an individual is willing to accept. We shall find convenient to refer to this as the reservation wage. If there is indifference to risk the risk premium is zero so the reservation wage is equal to the mean expected offer, but if there is risk aversion (positive p) job offers below the mean expected offer may be accepted, and if there is risk loving (negative p) only job offers that exceed the mean offer are accepted. We assume that a worker can have only one outstanding offer at a time and that the offers are written such that $Y_1 \leq Y_2 \leq \cdots \leq Y_n$.

If the distribution of wage offers is unchanging, differences in the risk premium will be due to differences in risk attitudes. Using Pratt's result that the degree of risk aversion varies directly with the risk premium and equation (6), we find that *the reservation wage varies inversely with the degree of risk aversion.*

Consider now two societies which are identical in all respects except the attitudes of their members toward risk. In particular we assume that both societies are in stochastic equilibrium in that there is a distribution of wage offers for similar types of labor and not a unique wage rate.[2] The number and type of firms making wage offers in the two societies are the same so the principles followed in determining wage offers in the two societies are the same, but individuals in the first society are assumed to be more risk averse than in the second. Thus the labor-supply function faced by the firms in the two societies is different.

Let Y_j^1 be the reservation wage of individual j in society 1, and Y_j^2 the reservation wage of the corresponding individual in the less risk-averse society 2. Since the two individuals are identical in all respects except their attitudes toward risk, the corresponding reservation wages in society 2 are greater than those in society 1. Thus firms in society 2 face a tighter labor market than those in society 1 in that if they offer the same wage rate their chances of finding an employee within a specified interval of time are less than those in society 1. A model of firm behavior under similar conditions, but ignoring risk attitudes, has been developed by Phelps (1970, pp. 131–42). He assumes that the typical firm chooses optimally a wage differential d^*, defined as the difference between the wage offer it makes and what it conceives to be the average wage in the market. The wage differential is a function of the factors which determine the chances of acceptance of a given offer (i.e., if a firm thinks that the chances of acceptance of a given wage offer have decreased, it puts up its wage differential to make the wage offer more attractive and thus restore the flow of employees). Phelps considers two such factors, the unemployment rate and the vacancy rate, arguing that a reduction in the former or an increase in the latter indicates a tightening of market conditions so the firm responds by increasing its wage differential. Extending Phelps's model to deal with risk, we may argue that the distribution of risk premia among individuals, $Q(p)$, will also affect the determination of the wage differential, in that if $Q(p)$ shifts to the right (increase in risk aversion) firms can more easily obtain employees at a given wage offer so they may put their wage differential down. Letting h denote an additive shift parameter in the distribution of risk premia, we may then postulate

---

[2] It is possible to prove the existence of such stochastic equilibria for markets with imperfect information when the participants follow a process of search, but the details are omitted here because they are not directly relevant to the arguments advanced. For a survey of formal stochastic equilibrium models, see Rothschild (1971). The reasons for the existence of a distribution of wages have been discussed extensively by Stigler (1962) and Phelps et al. (1970).

$$d^* = f[Q(p, h), \ldots] \qquad \partial f / \partial h < 0. \tag{7}$$

Letting now d_1^* denote the wage differential in the first society and d_2^* the same in the second society, it is clear that d_1^* is less than d_2^*, since we may set $h_1 > 0$ and $h_2 = 0$ without loss of generality (where h_i refers to society i, $i = 1, 2$). Thus the distribution of wage offers in the second society is to the right of the distribution in the more risk-averse society: *At any point of time the level of earned incomes will tend to be higher the lower the degree of risk aversion.*

Several of the results of the job-search literature may be modified to deal with risk along similar lines. In general, from the point of view of the firm the presence of risk preference is equivalent to a tightening of labor markets whereas risk aversion is equivalent to an easing of the market. Some strong results may also be obtained from a further reasonable restriction on the wage-setting equation (7). Thus, suppose that when h falls firms increase their wage offers, but not by the full amount of the shift in h (i.e. $\partial^2 f / \partial h^2 > 0$). This implies that, if prospective employees now will accept Y_j^0 instead of $Y_j^0 - h$ (positive h), the firm does not increase its wage offer by the full amount of h to restore fully the expected flow of labor supply, but adjusts to the tightening of the market by allocating the costs optimally between the money costs due to the increase in wage offers and the time costs due to an increase in the expected time it takes to find an employee. One of the implications of this assumption is that starting from any given point adjustment in the less risk-averse society will take longer than in the more risk-averse society, so in general, *the amount of search (frictional) unemployment will be lower the higher the degree of risk aversion.*[3]

We may proceed in a similar manner and obtain the other results of the job-search literature, such as the determinants of the rate of inflation and the slope of the Phillips curve, under alternative assumptions about risk. What has been said, however, should suffice as an indication of the effects of risk aversion, so we now turn to a related problem which has been largely neglected by the job-search literature but has been widely discussed in other contexts. The performance of the job-search model in confirming or contradicting the results of the income-distribution model under uncertainty should serve as a potential test of the theory, and vice versa.

---

[3] This result, of course, requires the crucial assumption that those unemployed are more flexible in undertaking search than those employed. See Alchian (1970) for statement and Tobin (1972) for criticism.

IV. The Distribution of Income and Wealth among Individuals

The distribution of income and wealth has been extensively discussed by economists, but most of the discussion concentrated on the distribution among factors of production and not individuals (see Bronfenbrenner [1971] for a discussion of the literature). The relationship between uncertainty, risk behavior, and the personal income distribution has been investigated by Friedman (1953), who showed that if individuals maximize expected utility and if in two societies the probability distribution of income offers is the same but risk attitudes are different, the one with the least risk aversion will show the greatest inequality in earned incomes. In this section we show that the model of Section III contradicts this result. In particular we show that risk aversion implies at least as great inequality of incomes as risk loving, and if in addition the risk premium declines as wealth increases, in the long run wealth inequalities will tend to be amplified through job search.

Suppose firms make offers sequentially until the jobs are accepted, and let the initial set of wage offers $Y^0 = (Y_1, \ldots, Y_n)$ be the same in both societies. Writing $Y^1_{\min}$ for the lowest reservation wage in society 1, and $Y^2_{\min}$ for the same in society 2, the incomes that are accepted in the first society must belong to the set $Y^1 = (Y^1_{\min}, \ldots, Y_n)$ and those accepted in the second society must belong to the set $Y^2 = (Y^2_{\min}, \ldots, Y_n)$. Since $Y^2_{\min}$ is greater than $Y^1_{\min}$ the range of incomes in the second society is already smaller than the range of incomes in the first and more risk-averse society. Now consider the second stage. Firms with vacancies in the second society face a tighter market than those in the first, so if an upward shift in wage offers occurs this shift cannot be greater in the first society than in the second. Also the difference in the shift between the second and first society will almost certainly be less than the difference between the reservation wages $Y^2_{\min} - Y^1_{\min}$. So now the set of offers in the second society extends beyond that of the first, but the set of acceptable incomes in the first society is still more spread out because the difference at the upper end is always below that of the reservation wages at the lower end.[4] One would expect incomes to be still more concentrated in the least risk-averse society because the tightness of the market is mainly felt by the firms which make the unattractive offers, so the incentive for upward revision of offers is greater at the lower end of the scale than at the higher. Thus one may witness revision of low offers in the least risk-averse society quicker than any other offers in the two societies, in this way pushing incomes in society 2 still closer together than they were initially. Therefore if firms follow similar wage policies in the two societies and individuals are also similar apart from their attitudes toward risk,

---

[4] Some revision of reservation wages may, of course, take place, but if individuals are alike short of risk attitudes, the revision in the two societies is not likely to be so divergent as to alter this result.

the range of incomes in the more risk-averse society is never smaller than the range of incomes in the less risk-averse society.

We now consider any one society in isolation and show that if the risk premium decreases when wealth increases (postulated by Arrow and Pratt to be reasonable risk-averse behavior) *wealth* inequalities will tend to be amplified in the long run if there is risk aversion. Let A_j denote the initial wealth of individual j and p_j the risk premium of the same individual, evaluated at $A_j + EY$. If two individuals, i and j, are identical in all respects except that A_i is greater than A_j, and if there is decreasing risk aversion, then p_i will be less than p_j, so the reservation wage of individual i will be higher than that of individual j. Thus the less wealthy individual is willing to accept a lower wage offer than the wealthier individual, and since both individuals are receiving offers from the same distribution individual j is more likely to find an acceptable offer than i, but individual i is likely to find a better-paying offer because he will spend more time searching. Thus in the long run the more wealthy individual is more likely to earn a higher income than the poorer one, so wealth inequalities will tend to be amplified.

V. Extensions of the Model

The results obtained in the last section are clearly counterintuitive and contradictory to existing results. Their derivation, however, depends crucially on the assumption that the distribution of wage offers is independent of risk attitudes, so it may be taken to be the same for both risk averters and risk lovers. This assumption, of course, is not needed for the individual analysis of Section III, where the distribution of offers is arbitrarily taken as given. But when generalizing to many individuals it will not in general be appropriate to take the distribution of wage offers as common, since job practice requires previous training and the choice of training will in general be influenced by risk attitudes.[5] A risk averter entering the market may have committed himself beforehand to a certain range of jobs by proper choice of training, and this range may be quite different from the range of jobs open to the risk lover. Risk averters may decide to train in "safe" jobs, while a risk lover may decide to train in "risky" occupations, such as entertainment, which nevertheless offer a few very high incomes. Under these circumstances when the risk averter joins the market to look for a job the variability of wage offers he receives will tend to be less than the variability faced by the risk lover. Thus the risk premium of the risk averter is not directly comparable with the risk premium of the risk lover in the same way as the risk-aversion coefficients are, because the risk premium depends on both the utility function and the distribution of wage offers whereas the risk-aversion coefficients are defined solely in terms of the utility function. When risk is taken into account, as it should be when there is imperfect knowledge, the assumption of a predetermined skill level is likely to contradict the assumption of a unique distribution of wage offers existing in the market. As a first approximation we may assume that the distribution of wage *offers* tends to be more concentrated as the degree of risk aversion increases.

---

[5] The Appendix formalizes some of the ideas that follow and shows how far the formal analysis supports the views expressed.

Consider a society consisting solely of risk averters, some of them more risk averse than others. Assume that all occupations offer the same mean wage but the variability of wages in some occupations is greater than the variability in others. Then all members of society will want to train in the safe occupations, wages in these jobs will subsequently tend to go down because of the excess supply and wages in the more risky jobs will tend to shift up, until enough people are attracted by them. Thus the mean wage in the more risky jobs will tend to be higher than the mean wage in the safe jobs, the difference being determined by the willingness of people to take the higher risk. Hence two causes of income inequality may be distinguished: first, the inequality between the mean incomes of safe and risky jobs determined by the willingness of people to bear risks, and second, the inequality of incomes within occupational groups, caused by labor-market imperfections. The difference between the mean income of safe and risky jobs will move from negative when there are only risk averters to positive when there are only risk lovers. If the number of risk averters and risk lovers is just enough to fill all safe and risky vacancies, respectively, then the means may be equal. Income inequality in this sense is greatest when all members of society take a similar attitude toward risk and when this attitude is at either extreme. When the individual chooses a job, his decision with respect to this case involves the choice of a distribution of wage offers, or the choice of an "occupation," from a large number of occupations (distributions).

Inequalities within occupational groups may be represented by the particular characteristics of the relevant distribution. Once an individual decides on his career he commits himself to a single distribution, so his decision now involves the choice of a point (a particular wage offer) from that distribution. It is this choice problem which is considered by the job-search literature and which was treated in Section III, but the distribution facing risk averters and risk lovers will be different because of the choice of career which the individual has already made. Risk lovers will normally choose a job offer out of a distribution which is less concentrated, so ultimately the distribution of incomes within a risk-loving group will tend to be more unequal than within a risk-averse group. Risk lovers will also tend to be unemployed longer because they will insist on the higher end of the income scale, but since only a few of these jobs are available they will eventually be forced to take less well-paid jobs. Whether risk lovers will on average be better paid than risk averters cannot be established because this will depend on the first kind of inequality, intergroup differences. Other things equal, in a society dominated by risk averters risk lovers will be better paid; and vice versa.

The conclusion to be drawn from this analysis is that the assumption of a unique distribution of wage offers in the market is unrealistic and may lead to paradoxical results. The work-choice problem may be broken up into two stages, (1) choice of career, and (2) choice of a particular job. The first choice involves the choice of a distribution of wages. the second the choice of a single wage. In the first stage, if all members of society have equal opportunity, they may be considered together because the constraint (the set of all careers) is common to all; in the second stage the distribution chosen by each occupational group becomes its own constraint, so each group must be considered separately. Income inequalities arising from the first stage refer only to *mean* differences, and they depend on the distribution of risk attitudes among the members of society and not on the degree of risk aversion. Inequalities arising from the second stage depend on the extent of labor-market imperfections and the degree of risk aversion, because risk averters will choose occupations with a well-developed information network and risk lovers occupations with less well-developed networks (and so greater variability of incomes). Thus on the whole a society consisting of risk lovers may exhibit greater income inequalities than a society consisting of risk averters, but there is no simple relationship between the degree of risk aversion and the extent of income and wealth inequalities.

VI. Remarks on the Empirical Relevance of the Analysis

In this paper we obtained some results about the relationship between risk behavior, job search, and the personal distribution of income. The analysis was strongly in the *ceteris paribus* tradition in that everything except risk attitudes was assumed constant. Since risk attitudes is just one factor in a whole chain of factors which are likely to influence man's decision with respect to such important instances as choice of career and job, empirical verification is almost impossible. To argue that because of this risk attitudes are irrelevant in practice would be equally wrong, so from the point of view of empirical studies of frictional unemployment and earned incomes the main contribution of the analysis is the isolation of a factor that is likely to introduce systematic biases in estimates if neglected. Such neglect, for instance, was one of the reasons that led Friedman and Kuznets (1945) to write, when noting that the ranking according to the mean level of income in the professions they examined corresponded to the ranking according to their variability, "We are inclined to regard this similarity among the rankings, striking though it is, as coincidental, since we can find no reasonable explanation for it" (p. 396). As it is clear from the analysis developed in Section V of this paper, this similarity in rankings is to be expected in a society dominated by risk averters. Similar conclusions as those of Friedman and Kuznets were reached by Miller (1955, chap. 3), also for U.S. data, and Lydall (1955, pp. 25–28) for British data. Miller attempted a rationalization of this correspondence along the lines suggested by Friedman (1953), but as Bronfenbrenner (1971, pp. 58–62) emphasized this is one of many explanations, and observations at this level of generality do no more than narrow the range of possible explanations, unfortunately not always to manageable proportions.

Empirical investigations, on the other hand, may be used to settle the issue where the theory offers no clear-cut conclusions, as in the extended model of Section V. There is at least one finding in the empirical literature which bears directly on the issues raised there, namely, Miller's (1955, chaps. 3, 5) finding that the distribution of earnings for broad occupational groups was largely symmetric, the right skewness in the total distribution arising because of the aggregation of group distributions with different means and variances. Thus if we apply the two-stage procedure proposed in Section V, a mean-variance approximation for the distribution of job offers may suffice, simplifying in this way the theoretical analysis and enabling empirical disentanglement of the income inequalities due to the nature of the distribution of risk attitudes among members of society, the extent of risk aversion, and the imperfections of the labor market. Without the aid of fairly strong restrictions, such as those implied by mean-variance rankings, such disentanglement in practice is impossible (see the Appendix).

Appendix

Let Y^0 denote the set of all wage offers and let Y^i $(i = 1, \ldots, I)$ be I subsets of Y^0 which may or may not have nonempty intersections. A cumulative probability distribution $P^i(Y)$ is defined over the elements of each Y^i. Each Y^i may be identified with a particular skill level (or "occupation"), so choice of training before entering the labor market and searching for a job may be interpreted as choice of a probability distribution $P^i(Y)$ out of all such distributions $P^1(Y), \ldots, P^I(Y)$. It is assumed that Y^0 is closed, continuous, and bounded from below by 0 and from above by 1. Training costs are assumed to be incorporated in each Y^i. An element Y of Y^i represents the present value of an individual if he trains in occupation i *minus* the costs of training for that occupation. Let $P^*(Y)$ denote the occupation chosen. Then if the individual obeys the von Neumann-Morgenstern axioms,

$$\int_0^1 U(A + Y)\, dP^*(Y) \geq \int_0^1 U(A + Y)\, dP^i(Y) \quad \text{for all } i = 1, \ldots, I.$$

Writing this as one integral and omitting the arguments of the functions for notational convenience, we obtain

$$\int_0^1 U(dP^* - dP^i) \geq 0.$$

Integrating this by parts gives

$$[P^*(1) - P^i(1)]U(A + 1) - [P^*(0) - P^i(0)]U(A) - \int_0^1 (P^* - P^i)\,dU \geq 0.$$

But $P^*(1) = P^i(1) = 1$, and we may let $U(A) = 0$, so this reduces to

$$\int_0^1 [P^*(Y) - P^i(Y)]\,dU(A + Y) \leq 0.$$

Thus if a probability distribution $P^*(Y)$ lies everywhere below all other distributions it is always chosen whatever the shape of the utility function (see also Hadar and Russell [1969], Hanoch and Levy [1969], and Rothschild and Stiglitz [1970]). This is the only case where the distribution of expected wage offers is independent of risk attitudes. It is a very restrictive case, and even if it arises in practice it is unlikely to persist because the increased supply of labor to that occupation will eventually drive the probabilities of receiving a given offer down. Once two distributions meet at any point, choice between them depends on the *shape of the utility function.*

Assume now that an individual is either risk averse or risk loving but does not show features of both at different levels of wealth (i.e., the utility function is either concave or convex everywhere). Then it can be shown (see the papers referred to in the previous paragraph) that risk averters will choose distribution $P^*(Y)$ if and only if

$$\int_0^x [P^i(Y) - P^*(Y)]\,dY \geq 0 \qquad \text{for all } 0 \leq x < 1; \quad i = 1, \ldots, I.$$

Now this condition imposes only a partial ordering on the set of all occupations, since if the integral changes sign at some level of wealth x some risk averters may choose $P^*(Y)$ and some others $P^i(Y)$. Unfortunately it is not possible to deduce a relationship between the distribution chosen and a measure of risk aversion, such as the absolute risk-aversion coefficient, so if the class of distributions is not restricted to the class which satisfies the above integral condition it is not possible to derive the relationship between the distribution of wage offers and the distribution of risk attitudes without restricting the utility function.

On the other hand if we restrict the utility function (e.g., to be quadratic) or use mean-variance analysis as an approximation, it *is* possible to deduce a relationship between risk attitudes and the distribution of offers because then the risk premium becomes (see Pratt 1964) $p^i = -\frac{1}{2}RV^i(Y)$, where R is the absolute risk-aversion coefficient evaluated at $A + EY$ and $V^i(Y)$ is the variance of $P^i(Y)$. The mean and variance of the distribution order all $P^i(Y)$ completely, so given that an individual will choose a distribution with highest certainty-equivalent $E^i(Y) - p^i$, we may write: $P^*(Y)$ is chosen if and only if $E^*(Y) - \frac{1}{2}RV^*(Y) \geq E^i(Y) - \frac{1}{2}RV^i(Y)$ for all i. By varying R we obtain different preferred distributions, so we deduce a well-defined relationship between a measure of risk attitudes, R, and the first two moments of the distribution of offers, the difference between $E^*(Y)$ and $E^i(Y)$ measuring between-group inequalities and $V^*(Y)$ measuring within-group inequalities.

References

Alchian, A. A. "Information Costs, Pricing, and Resource Unemployment." In *The Microeconomic Foundations of Employment and Inflation Theory*, by E. S. Phelps et al. New York: Norton, 1970.

Arrow, K. J. *Aspects of the Theory of Risk-bearing*. Helsinki: Yrjo Jahnssoninsaatio, 1965.

Bronfenbrenner, M. *Income Distribution Theory*. London: Macmillan, 1971.

Friedman, M. "Choice, Chance, and the Personal Distribution of Income." *J.P.E.* 61, no. 4 (August 1953): 277–90.

Friedman, M., and Kuznets, S. *Income from Independent Professional Practice*. New York: Nat. Bur. Econ. Res., 1945.

Hadar, J., and Russell, W. R. "Rules for Ordering Uncertain Prospects." *A.E.R.* 59 (1969): 25–34.

Hanoch, G., and Levy, H. "The Efficiency Analysis of Choices Involving Risk." *Rev. Econ. Studies* 36 (1969): 335–46.

Hicks, J. R. *Value and Capital*. 2d ed. Oxford: Clarendon, 1946.

Hutt, W. H. *The Theory of Idle Resources*. London: Jonathan Cape, 1939.

Keynes, J. M. *A Treatise on Money*. Vol. 2. London: Macmillan, 1930.

Lydall, H. F. *British Incomes and Savings*. Oxford: Blackwell, 1955.

Miller, H. P. *Income of the American People*. New York: Wiley, 1955.

Phelps, E. S. "The New Microeconomics in Employment and Inflation Theory." *A.E.R. Papers Proc.* 59 (1969): 147–60.

———. "Money Wage Dynamics and Labor Market Equilibrium." In *The Microeconomic Foundations of Employment and Inflation Theory*, by E. S. Phelps. New York: Norton, 1970.

Phelps, E. S., et al. *The Microeconomic Foundations of Employment and Inflation Theory*. New York: Norton, 1970.

Phillips, A. W. "The Relation between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1862–1957." *Economica* 25 (1958): 283–99.

Pratt, J. W. "Risk Aversion in the Small and in the Large." *Econometrica* 32 (1964): 122–36.

Rothschild, M. "Models of Market Organization with Imperfect Information: A Survey." Paper presented at the December 1971 meetings of the Econometric Society.

Rothschild, M., and Stiglitz, J. E. "Increasing Risk. I. A Definition." *J. Econ. Theory* 2 (1970): 225–43.

Stigler, G. J. "Information in the Labor Market." *J.P.E.* 70, no. 5, pt. 2 (October 1962): 94–105.

Tobin, J., "Inflation and Unemployment." *A.E.R.* 62 (1972): 1–18.

von Neumann, J., and Morgenstern, O. *The Theory of Games and Economic Behavior*. 2d ed. Princeton, N.J.: Princeton Univ. Press, 1947.

The Role of Absolute and Relative Wage Differentials in the Choice of a Work Activity

Shmuel Sharir and Yoram Weiss

Tel-Aviv University

A basic economic hypothesis states that, other things being equal, an increase in the wage at a certain work activity will attract labor to it. Economic theory tells us that the list of "other things" includes prices of goods, wages in alternative activities, and real income. But what if other things are not equal and, in particular, if wages in alternative work activities vary simultaneously? It then seems natural to concentrate on wage *differentials*. It is said that labor will be attracted to the types of employment where wages increased more. There are, however, several questions concerning the precise nature of this approach: Under what conditions is it possible to dispense of the wage *levels*? That is, what are the conditions under which wage differentials play an *independent* role, and what are in these cases the "other" factors which affect supply? Finally, should wage differentials be measured in absolute or relative terms?

The purpose of this note is to analyze the role of absolute and relative wage differences in *short-run*[1] decisions concerning the choice of work activities within a given occupation and level of skill. Our framework is static demand theory. We focus on individual supply functions and assume that the individual views the amounts of time which he works (or may work) at different jobs as different commodities. Initially, we assume that it is feasible to combine work activities. The essentials of this approach have been outlined by Becker (1965, 1971, pp. 166–70).

The first section of the paper is concerned with alternative specifications of the supply functions and the role of relative and absolute wage differentials in these specifications. We note that even if income (properly defined) is held constant, neither absolute nor relative wage differentials are sufficient to determine the allocation of labor. Either the wage level

---

Helpful comments were received from Yoram Ben-Porath, Jacob Paroush, and Richard Perlman. The study was financially supported by the Foerder Institute.

[1] The role of absolute and relative wage differences in *long-run* choices, that is, educational and occupational choices, has been discussed by Becker (1964, pp. 52–53) and Freeman (1971, chap. 1).

[*Journal of Political Economy*, 1974, vol. 82, no. 6]

© 1974 by The University of Chicago. All rights reserved.

(reflecting the price of leisure and other nonwage activities) must be specified, or, alternatively, some additional constraints must be imposed on the allocation of time. A possible constraint is to fix the total amount of work. In this case the allocation of time to the various work activities is determined by *absolute* wage differentials. Another type of constraint which may be considered is on the total *value* of work. For instance, labor earnings or consumption may be predetermined. Such constraints lead to the relevance of *relative* wage differentials. Both types of constraints on the allocation of time seem to be of empirical significance.[2] For the sake of brevity, however, we shall concentrate on illustrating the effects of the constraint on the total *amount* of work.

In the second section, we investigate the effects of simultaneous compensated wage changes on the allocation of labor. Our purpose is to predict these effects on an a priori level. Again some additional constraints on the allocation of time are required. Specifically, it is necessary to hold some *quantity* index of total work constant. If the simple sum of the time inputs in the various work activities is predetermined, then the *absolute* wage differentials are the appropriate measure of attraction. If a weighted sum (where a set of fixed real wages serve as weights) is held constant, then *relative* wage differentials become relevant.

The third section of the paper contains a brief discussion of the special case in which combining of time activities is not feasible.

I

Let there be n time activities. Some of these are work activities with a positive wage; others are leisure activities which provide zero wage.[3] Let t_i be the proportion of the period spent at activity i and w_i the *real* wage per period in that activity, where $w_i \geq 0$. The individual's preferences are described by a utility function which depends on the bundle of goods he consumes and on the allocation of time among the various activities. To simplify, we assume that the relative prices of goods are fixed so that they can be aggregated into one composite good which we denote by x. The marginal utility of x is assumed to be positive. Nothing is assumed on the marginal utilities of the various time activities, and they may be positive or negative. Our only hypothesis is that the generalized law of decreasing marginal utilities holds and that the utility function is thus quasi-concave.

---

[2] Uniform standard working week and transfer costs may restrict the individual's total amount of work. The spread of installment credit and interdependencies in utility (see Duesenberry [1967, pp. 28–32]) may cause individuals to determine their earnings, at least in the short run, according to prior set consumption or earning targets. Behavior according to an earning target is said to motivate moonlighting (see Wilensky 1963; Mott 1965) and is presumably behind some overtime work.

[3] One could generalize the discussion by incorporating negative wages. A working woman, for instance, may in fact pay for the presumably preferred opportunity to work in the market. The payment is in terms of a possibly positive difference between commuting costs, maid's salary, etc., and her market wage. The introduction of negative wages has no effect on our conclusions.

The individual's maximization problem can be formulated as

$$\max_{x, t_1, t_2, \ldots, t_n} \quad U(x, t_1, t_2, \ldots, t_n) \tag{1}$$

subject to the income and time constraints

$$x = \sum_{i=1}^{n} w_i t_i + A, \tag{2}$$

$$\sum_{i=1}^{n} t_i = 1, \qquad t_i \geq 0, \tag{3}$$

where

$$\sum_{i=1}^{n} w_i t_i$$

is total real labor earnings and A is real property earnings during the period under discussion. We use x as a numeraire, and its price is therefore unity. The first-order conditions for interior[4] maximum are:

$$u_{t_i} + u_x w_i - \mu = 0, \qquad i = 1, 2, \ldots, n, \tag{4}$$

where μ is the shadow price of time.

The supply (demand) functions for the various time activities can be derived from the solution of equations (2), (3), and (4). Due to the assumed quasi-concavity of the utility function, this solution is unique.

In general, supply for each work activity will depend on all real wages, each one separately, and on real nonwage income. One can transform this relationship to include wage differentials and an appropriate income measure. But in this transformed formulation one wage *level* will still appear, so that wage differentials have no independent role. The wage level enters the supply relations in two distinct roles: one is obviously in its effect on the level of real income; the other we may term the price-of-leisure effect. As long as there are activities in which the wage is identically zero—that is, nonmarketed (leisure) activities—their relative price will be affected by a uniform increase in all work (market) activities. Conceptually there is an important difference between the two effects. The income effect may be eliminated (compensated) by an appropriate change in nonwage income. The same is not true with respect to the price-of-leisure effect, which will be present as long as *total* supply of

---

[4] The possibility of corner solutions in the choice of time activities should not be ignored. The more general statement of the first-order conditions is:

$$\begin{aligned} &u_{t_i} + u_x w_i - \mu = 0 \quad \text{if } 0 < t_i < 1, \quad u_{t_i} + u_x w_i - \mu \leq 0 \quad \text{if } t_i = 0, \\ &u_{t_i} + u_x w_i - \mu \geq 0 \quad \text{if } t_i = 1 \quad \text{for } i = 1, 2, \ldots, n. \end{aligned} \tag{4'}$$

labor (and not merely its allocation) is subject to choice. Only when total supply of labor is predetermined are wage differentials and an appropriate measure of income sufficient to determine the allocation of labor.

We shall now show that if the amount of work is predetermined—that is,

$$\sum_{i \in I} t_i = T^o,$$

where I is the set of all work activities, ($w_i > 0$ for $i \in I$)—then the allocation of work depends on *absolute* wage differences. In this case the first-order conditions (4) may be broken into

$$u_{t_i} + u_x w_i - \mu_1 = 0 \qquad \text{for} \qquad i \in I \tag{5}$$

and

$$u_{t_i} - \mu_2 = 0 \qquad \text{for} \qquad i \notin I,$$

where μ_1 and μ_2 are the shadow prices of work and leisure activities, respectively.

Let x^o, t^o be a solution of (2), (3), and (5), for a given vector w_1^o, $w_2^o, \ldots, w_m^o$ of wages and given nonwage income A^o. It can be seen from inspection of the constraints and first-order conditions that x^o, t^o will also be a solution when wages change to $w_1^o + \Delta w$, $w_2^o + \Delta w, \ldots,$ $w_m^o + \Delta w$ and nonwage income is reduced to $A^o - t^o \Delta w$. (Because of the uniqueness assumption, it is sufficient to show that there are admissible μ_1 and μ_2 which solve the problem with the original x^o, t^o and the *new* wage vector.) It follows that the supply (demand) functions for the various time activities may be written as

$$t_i^s = s_i(w_1 - w_m, w_2 - w_m, \ldots, w_{m-1} - w_m, A + T^o w_m, T^o),$$
$$i = 1, 2, \ldots, n, \tag{6}$$

where by assumption $w_i > 0$ for $i = 1, 2, \ldots, m$ and $w_i = 0$ for $i = m + 1, m + 2, \ldots, n$. If w_m is the maximal wage level, we may view $A + T^o w_m$ as the individual's "full" income (see Becker 1965). Equation (6) states that the allocation of labor depends upon (real) absolute wage differences, full income, and the predetermined amount of work.[5] Note

---

[5] It is important to note that the total supply of labor is an aggregate which may be defined in various ways. Instead of simple addition of time activities, we could define labor supply as

$$V = \sum_{i \in I} w_i^o t_i$$

where an appropriate set of *fixed* (real) wages serve as weights (see Friedman 1967, p. 202). Under this definition, the supply for each time activity may be written as

$$t_i^s = s_i[w_1/w_m, w_2/w_m, \ldots, w_{m-1}/w_m, (x - A)/w_m, V^0], \qquad i = 1, 2, \ldots, n \tag{6'}$$

$w_i > 0$ for $i = 1, 2, \ldots, m$, $w_i = 0$ for $i = m + 1, \ldots, n$. That is, given the value of supply of labor, V^0, the allocation of work is determined by *relative* wage differentials and the real costs (in terms of time) of achieving the original level of consumption. Indeed, in that given the level of full income, the wage *level* as such is of no relevance.

It should be emphasized that the appearance of *absolute* wage differentials in the supply function (6) is a direct result of the constraint on the total *amount* of work. Other constraints will lead to different results. In particular, if the individual has a prescribed (and presumably feasible) consumption (earnings) target, then supply to the various work activities is completely determined by *relative* wage differentials, the consumption target x^*, and the labor costs $(x^* - A)/w_m$ which are necessary to attain the consumption target.[6]

II

The analysis of the previous section concentrated on the correct specification of the supply relations. We now wish to examine the signs of the partial derivatives of the supply functions. As is well known, the effects of uncompensated wages changes, even solitary ones, cannot be predicted a priori. We shall focus, therefore, on *compensated* wage changes.

Let the wages in two jobs, say 1 and 2, change simultaneously while all other wages remain unchanged. Specifically, let $\Delta w_1 > \Delta w_2 > 0$. Assume further that nonwage income is adjusted so that real income (in terms of the original bundle) is unchanged.[7] Standard demand theory tells us that the individual's adjustment to this change is given by the "generalized substitution effect." Specifically, we have [8]

$$\Delta w_1 \Delta t_1 + \Delta w_2 \Delta t_2 \geq 0. \tag{7}$$

---

his discussion of occupational choice, Friedman (1967, pp. 211–13) presents the allocation of time between two work activities ("occupations") as a function of the relative wage differential. Becker (1971, pp. 167–70) uses the absolute wage differential for just the same purpose. The difference seems to be only methodological, and it apparently arises from two alternative definitions of the *total* supply of labor, which is (implicitly) held constant.

[6] Let x be exogenous. The first-order conditions for maximum become

$$u_{t_i} + \psi w_i - \mu = 0, \qquad i = 1, 2, \ldots, n, \tag{4''}$$

where ψ is the shadow price of the consumption goal x^*. (For simplicity, we assume the solution to be interior.) Let t^o be the solution of (4″) and the constraints, for a given x^*, A^o and w^o. We assume that x^* is set so that such a solution is feasible. Let wages change to γw^o, and let nonwage income change to $x^* - \gamma(x^* - A^o)$. It is seen that there is a pair ψ, μ which, together with t^o, solves (2), (3), and (4″). (Specifically, $\psi_1 = (1/\gamma)\psi_o$ and $\mu_1 = \mu_o$. It follows that the supply for the various work activities can be written as

$$t_i^s = s_i[w_1/w_m, w_2/w_m, \ldots, w_{m-1}/w_m, (x^* - A)/w_m, x^*], \qquad i = 1, 2, \ldots, n. \tag{6''}$$

[7] The costs of a given bundle x^o, t^o may be defined as $x^o - wt^o$. If real wages change from w^o to w^1, then the costs of x^o, t^o change by $-t^o(w^1 - w^o)$.

[8] The generalized substitution effect is a direct result of the assumption that, given the constraints on x and t, the minimum cost solution for a specified utility level is unique. Sufficient conditions for this purpose are strict quasi-concavity of the utility function and convexity of the feasible set for t and x. Note that these requirements are satisfied in the cases considered in the text.

It is immediately seen that if the total amount of time spent at the two activities, $t_1 + t_2$, is constant, then we obtain from (7)

$$\Delta t_1(\Delta w_1 - \Delta w_2) > 0.$$

That is, a simultaneous compensated wage increase will attract labor to the work activity in which the *absolute* increase is higher.[9]

Jobs 1 and 2 in the above discussion may in fact be aggregates (simple sums) of work activities, provided that *absolute* wage differences in each group remain unchanged. Thus, for instance, if wages in *all* work activities except in activity 1 increase by the same magnitude when total working time is fixed, then (after compensation) labor will be attracted to job 1 if and only if the absolute increase in w_1 exceeds the absolute increase in all other wages.

If more than two wages change simultaneously and by different magnitudes, the restriction of fixed amount of time is insufficient to determine the effects of compensated wage changes. In particular, $\Delta w_1 > \Delta w_2 > \Delta w_3 > 0$ does not imply $\Delta t_1 > \Delta t_2 > \Delta t_3$.

III

It is quite possible that because of transfer costs and fixed standard working periods, the combining of work activities is not feasible and the individual is restricted to work at only *one* job. There seems to be no reason why this restriction should affect our results. Nevertheless, it may be of interest to examine this apparently more realistic case of exclusive alternatives in some detail.

For simplicity, we assume that there is only one leisure activity, say activity n. It then follows that for a given level of (real) nonwage income, and a given (real) wage vector, there is a *unique* weak ordering of the

---

[9] Again, if a different aggregation rule is used, different results will be obtained. Suppose that either the Laspeyers or Paasche quantity measure of labor supply remains constant; then we obtain from (7)

$$\Delta t_1\left(\frac{\Delta w_1}{w_1^0} - \frac{\Delta w_2}{w_2^0}\right) > 0 \qquad \text{when } w_1^0 t_1 + w_2^0 t_2 = w_1^0 t_1^0 + w_2^0 t_2^0,$$

and

$$\Delta t_1\left(\frac{\Delta w_1}{w_1^1} - \frac{\Delta w_2}{w_2^1}\right) > 0 \qquad \text{when } w_1^1 t_1 + w_2^1 t_2 = w_1^1 t_1^0 + w_2^1 t_2^0.$$

In this case it is *relative* wage differentials which "matter," while absolute differences are irrelevant. Specifically, suppose that there are only two work activities available to the individual and that consumption is predetermined then for compensated wage changes the Paasche quantity measure remains constant, and it is the *relative* wage differentials which are relevant. It was pointed out to us by the referee that these results bring to mind a proposition by Houthakker (1952, p. 159), who shows that if two prices change simultaneously in a compensating fashion so that the Laspeyers price index is held constant, then it is possible to predict *relative* changes in quantities bought.

alternative jobs, which is given by

$$u_i = u(A + w_i t_i, 0, 0, \ldots, 0, t_i, 0, \ldots, 0, 1 - t_i), \qquad i = 1, 2, \ldots, n - 1. \quad (8)$$

Suppose that there is a *uniform* standard working period so that either $t_i = T^o$ or $t_i = 0$. It is immediately seen from (8) that the ordering of the alternative jobs is unaffected by an increase in all wages by the same magnitude, Δw, and a decrease in A by $T^o \Delta w$. The supply function (6) is therefore still valid, and the allocation of work is affected by *absolute* wage differentials.

Consider now a simultaneous *compensated* wage change. Suppose the individual originally worked at job i. To compensate him we put $\Delta A = -T^o \Delta w_i$. It is clear that if $\Delta w_i > \Delta w_j$, then, after compensation, potential consumption (and thus utility) at job j is reduced. If this holds for all $j \neq i$, the individual will stay at job i. If $\Delta w_j > \Delta w_i$ for some j, the individual may switch to job j. After compensation, his *propensity* to choose j will increase. In this sense, the effects of changes in *absolute* wage differentials can be predicted on an a priori basis when supply is predetermined in physical time units.

As in the previous case in which the combining of work activities was feasible, the appearance of *absolute* wage differentials reflects the constraint on the total *amount* of work. In the case of exclusive alternatives, this constraint is reduced to the requirement that at each job the individual works the same fixed amount of time. If the amount of work at each job were variable but restricted so that *earnings* were equal across jobs, then *relative* wage differentials would have been relevant. In the absence of any restriction on the amount of work in the different jobs, wage differentials are insufficient to determine the allocation of time, and knowledge of the actual wage *level* at each job is necessary.

References

Becker, Gary S. *Human Capital.* New York: Nat. Bur. Econ. Res., 1964.

———. "A Theory of Allocation of Time." *Econ. J.* 75 (September 1965): 493–517.

———. *Economic Theory.* New York: Knopf, 1971.

Duesenberry, James S. *Income, Savings, and the Theory of Consumers' Behavior.* New York: Oxford Univ. Press, 1967.

Freeman, Richard B. *The Market for College-trained Manpower.* Cambridge, Mass.: Harvard Univ. Press, 1971.

Friedman, Milton. *Price Theory.* Chicago: Aldine, 1967.

Houthakker, H. S. "Compensated Changes in Quantities and Qualities Consumed." *Rev. Econ. Studies* 19, no. 3 (1952): 155–64.

Mott, Paul E. "Hours of Work and Moonlighting." In *Hours of Work,* edited by Clyde E. Dankert, Floyd C. Mann, and Herbert R. Northrup. New York: Harper & Row, 1965.

Wilensky, Harold L. "The Moonlighter: A Product of Relative Deprivation." *Indus. Relations* 3 (October 1963): 105–24.

Stability in the Pareto Law and the Popular Folklore: A Comment

Daniel R. Vining, Jr.

Carnegie-Mellon University

In a recent paper appearing in this *Journal*, Ijiri and Simon (1971) contrast the conventional wisdom of an ever-increasing concentration of economic power with the remarkable stability in the slope parameter of the Pareto law, β. The explanation they put forward to account for this discrepancy between the data and what they term the "popular folklore" is that the public has confused an increase in merger activity with an increase in business concentration. By demonstrating that there is no necessary connection between the two and that in fact there are plausible conditions under which an increase in merger activity will leave β unaltered,[1] their article might be construed as serving to dispel an illusion that has long existed in the public mind. To dispel illusion and superstition where it exists is a valuable and legitimate function for economists to perform, and the Ijiri-Simon paper deserves our attention for that reason alone.

The purpose of this note is to show that, in fact, there is no contradiction between the "popular folklore" and a stable β parameter, simply because the properties of the economic system captured by the β parameter are not likely to be of interest to those who maintain the popular view. Rather, it is the behavior of the other parameter of the Pareto law, consistently omitted from discussion in their papers, that is most probably its basis. It will be remembered that the Pareto law is in fact a two-parameter law (see Parzen 1960, p. 211): one, β, giving the slope of the log-linear relationship between rank and size and the other, x_0, the lowest size for which the relationship holds, that is, the one below which it is nonlinear. It is entirely conceivable that the former is stable over time and the latter highly unstable. In particular, the latter may be increasing, in which case some firms which were before of sufficient size to grow in accordance with Gibrat's law will find themselves of insufficient size to do so and there is a corresponding shift in relative strength to those firms

---

[1] Or relatively unaltered. In a more recent paper, using better and more recent data, Ijiri and Simon (1974) find that β would have been 7 percent lower had no mergers taken place over the last 20 years.

[*Journal of Political Economy*, 1974, vol. 82, no. 6]

© 1974 by The University of Chicago. All rights reserved.

still above the cutoff point, or what Ijiri and Simon have called elsewhere "minimum efficient size" (see Ijiri and Simon 1964, p. 81). To many observers, such a development would constitute prima facie evidence of an increasing degree of business concentration.[2]

There is good evidence, in the United States at least, that the two parameters have behaved in precisely this fashion. One authority on firm size distributions, for example, reports the following (Steindl 1969, pp. 187–89):

> The comparative stability in the distribution is illustrated by the U.S. corporations. . . . Between 1931 and 1955 the slope declines, reflecting the increasing concentration of business, but the change is small for the long period concerned. Again, the patterns of distribution of U.S. firms according to number of employees, as judged by mere visual impression, are very similar in 1939 and 1956.
>
> *An important peculiarity may be noted, however, in the case of U.S. corporations.* As already noted, only less than 2 percent of the corporations in 1955 were within the range of the strict Pareto law (straight line), but in 1931 more than 10 percent were in this range. Intuitively speaking, this appears to be due to the concentration of corporate business in the intervening period: *a greater inequality is expressed in a shift of the point where the curvature of the distribution ends and it becomes a straight line. The Pareto coefficient itself reflects this increased inequality only mildly: it decreased just a little in the period concerned.* [Emphasis added]

That is, while inequality among the population of firms above the minimum efficient size has remained relatively stable, this population itself has grown progressively smaller as a percentage of all firms. It is this latter phenomenon that persons most commonly will have in mind when they speak of a progressive concentration of economic power.

Another way of seeing this is to consider the percentage of the economy's total assets possessed by the corporations studied by Ijiri and Simon (i.e., the *Fortune 500*). If the cutoff point, x_0, had been stable over a long period of time, then this percentage should likewise be stable, for with a stable cutoff point the aggregate assets of those below the cutoff will

---

[2] A stable β parameter alongside a rising cutoff point, x_0, is entirely consistent with the Ijiri-Simon model. As is demonstrated in the early articles by Simon (see Simon 1955 and Simon and Bonini 1958), the former is a function of the rate of entry of new firms into the population of firms growing in accordance with Gibrat's law and is therefore quite conceivably independent of the magnitude and rate of change of the minimum size, x_0, at which they enter this population. Indeed, in the actual mathematics of the model, the cutoff point is essentially normalized out of existence by always being assumed to be one (see Ijiri and Simon 1964, p. 81); this procedure is tantamount to assuming that they are independent.

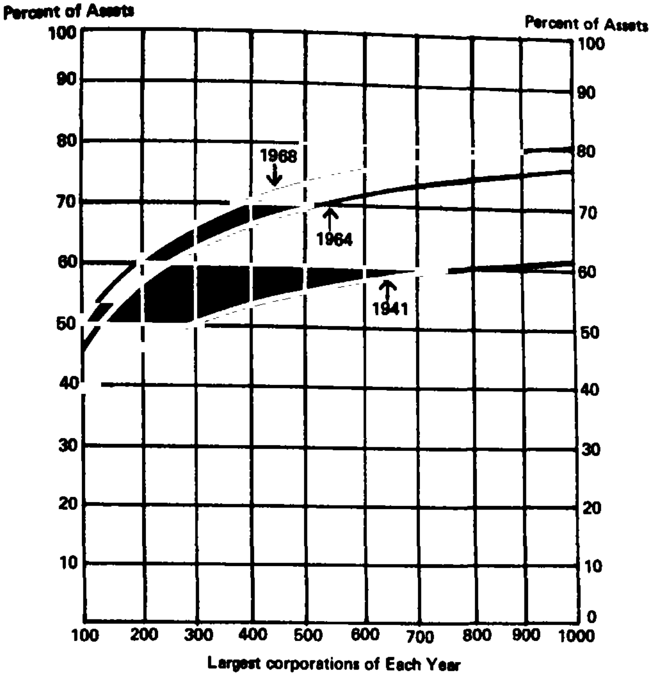


FIG. 1.—Cumulative share of corporate manufacturing assets held by 1,000 largest corporations: 1941, 1964, and 1968. From *The Industry and Enterprise Structure of the U.S. Economy* by Lee E. Preston. © 1971 General Learning Corp. Reprinted by permission.

diminish in importance, leaving nothing for those above to grow at the expense of.[3] In fact, this percentage has increased from slightly more than 55 percent in 1941 to 75 percent in 1968 (see fig. 1). This suggests that

---

[3] In fact, if all firms in the economy were growing in accordance with Gibrat's law, then the market share of the 500 largest of those firms should be declining over time. Consider a population of firms with a total aggregate size of N. Let n be the aggregate size of the 500 largest of those firms. Their share of total assets is, accordingly, n/N. Now, consider the next asset to be added to this population. The probability of its being added to one of the 500 largest is

$$(1 - \alpha)n/N, \tag{1}$$

where α is the probability that the new asset is added to a new firm at the minimum size, 1. Thus, the expected size of the 500 largest, after the addition of the new asset, is

$$n + (1 - \alpha)n/N. \tag{2}$$

And their expected share of total assets is then

$$[n + (1 - \alpha)n/N]/(N + 1). \tag{3}$$

Clearly, (3) is less than their share previous to the addition of the new asset, n/N. Thus, their expected share is ever declining over time. In Steindl (1965, p. 72), we find the opposite conclusion, namely, that their share will grow continuously in a growing economy. I am unable to fathom the reasoning that led to this result.

the cutoff point has risen rapidly over this period, causing a massive shift of relative strength from firms newly fallen below the cutoff point to those remaining above.

Figure 1 summarizes the point of this note rather well. While the form and curvature of the less-than cumulative have remained stable over time, its location has shifted upward dramatically. The Ijiri-Simon model is an attempt to reconcile the former fact with that of heightened merger activity. The latter development lies outside their analysis, although it doubtless is the source of the belief on the part of a large class of this country's citizenry that the concentration of economic power is on the increase.

References

Ijiri, Y., and Simon, H. A. "Business Firm Growth and Size." *A.E.R.* 54 (March 1964): 77–89.

———. "Effects of Mergers and Acquisitions on Business Firm Concentration." *J.P.E.* 79 (March/April 1971): 314–22.

———. "Interpretations of Departures from the Pareto Curve in Firm-Size Distributions," *J.P.E.* 82, no. 2, pt. 1 (March/April 1974): 315–32.

Parzen, E. *Modern Probability Theory and Its Applications.* New York: Wiley, 1960.

Preston, L. E. *The Industry and Enterprise Structure of the U.S. Economy.* New York: General Learning Press, 1971.

Simon, H. A. "On a Class of Skew Distribution Functions." *Biometrika* 42 (December 1955): 425–40.

Simon, H. A., and Bonini, C. P. "The Size Distribution of Business Firms." *A.E.R.* 48 (September 1958): 607–17.

Steindl, J. *Random Processes and the Growth of Firms: A Study of the Pareto Law.* New York: Hafner, 1965.

The Real Stock of Money and Velocity in a Wartime Economy: A Comment

Joong-Koon Lee

University of Cincinnati

How rapidly does the public adjust its price expectations during wartime?[1] Many writers, including M. Friedman and some of his former students (1952, 1956), have studied the pace of adjustment of price expectations in wartime economies, as revealed by changes in the ex post behavior of the real stock of money and various measures of velocity. The article by G. E. Makinen, recently published in this *Journal* (1971), in which he examined the Korean data for the wartime years, 1950–53, and the Vietnamese data for more recent years, can be considered as another attempt to shed more light on this issue and should be welcome.

It is my judgment, however, that Makinen's findings (pp. 1237–42) regarding drastic changes in Korean "expenditure" velocity between 1949 and 1951 are incorrect.[2] Makinen computed a 44 percent reduction in velocity during 1950 (table 14, p. 1237). Invoking Cagan's two-lag hypothesis to explain this drop and subsequent rise in velocity, Makinen claims that there was a long lag in the adjustment of price expectations as evidenced by actual balances running ahead of desired balances in 1950. It is my contention that lags of such magnitudes were not evident in the Korean experience, and Makinen's conclusion resulted from a faulty handling of statistical information.

When examined within a simple equation of exchange framework, it will be discovered that Makinen's estimated value for velocity is simply inconsistent with known values of the other terms in the equation. A further investigation makes it quite obvious that Makinen used an unduly low estimate for 1950 GNP and an unduly high value for 1949 imports of goods and services.[3] It is known that real output declined by 15.1

---

[1] I wish to thank Joseph L. Craycraft, Lloyd Valentine, and Harland W. Whitmore for their helpful comments.

[2] Although discrepancies exist between our calculations and Makinen's for particular yearly values of wartime velocity in Vietnam, the relative movements from year to year appear to agree with his findings. Hence, his results for Vietnam are not under question in this paper.

[3] As far as we can determine, no estimate of GNP in current prices has been published for the period 1949–52. All that is available are the index of total real output and various price indices.

[*Journal of Political Economy*, 1974, vol. 82, no. 6]
© 1974 by The University of Chicago. All rights reserved.

percent in 1949–50 and by 6.1 percent in 1950–51 (Bank of Korea 1965, p. 20), and that the wholesale price index rose by 140 percent in 1949–50 and by 262 percent in 1950–51 (Bank of Korea 1964, p. 23).[4] It is possible, therefore, to estimate the probable range of the 1950 GNP in current prices on the basis of the 1949 GNP of 12.6 billion hwan and the 1951 value of 102.0 billion hwan, provided by Makinen in his table 14:

$$\begin{aligned}\mathrm{GNP}_{(1950)} &= \mathrm{GNP}_{(1949)} \cdot \frac{y_{(1950)}}{y_{(1949)}} \cdot \frac{p_{(1950)}}{p_{(1949)}} \\ &= (12.6) \cdot (0.849) \cdot (2.404) = 25.7 \text{ billion hwan}\end{aligned}$$

$$\begin{aligned}\mathrm{GNP}_{(1950)} &= \mathrm{GNP}_{(1951)} \cdot \frac{y_{(1950)}}{y_{(1951)}} \cdot \frac{p_{(1950)}}{p_{(1951)}} \\ &= (102.0) \cdot (1.065) \cdot (0.276) = 30.0 \text{ billion hwan.}\end{aligned}$$

Hence, Makinen's figure of 16.3 billion hwan appears to understate the 1950 GNP by as much as 9–14 billion hwan.

It is also impossible to duplicate Makinen's estimate of 3.1 billion hwan for imports in 1949. If his value is correct, it means that (commercial plus aid-financed) imports reached a level of 25 percent of total output—an impossibly high figure for 1949.[5] However, according to data we have accumulated, commercial imports in 1949 amounted to 147 million hwan (or $16.3 million) and the aid-financed imports a total of $116.5 million (Bank of Korea 1956, p. 9). Converting the latter dollar value into an equivalent hwan measure by applying the official exchange rate of 9.0 hwan per dollar, as Makinen did,[6] the aid-financed imports amounted to 1,049 million hwan. Therefore, the imports for the year totaled 1,196 million (or $132.8 million), a figure that is not even remotely close to 3.1 billion hwan reported by Makinen. Given that exports were 113 million hwan for the year (ibid.), the net imports amounted to no more than 1.08 billion hwan in 1949.

---

[4] The percentage changes in the wholesale price index are computed on the basis of annual averages of June and December price indices only. Although the averages of monthly or quarterly figures would be preferable, such averages could not be used due to the fact that the relevant figures for certain months during the active phase of the war are missing from the published statistical tables.

[5] That Makinen's figure for imports in 1949 is an unduly high estimate can also be illustrated by reference to the Nathan report (1954). According to this report (p. 88), net imports for the fiscal year 1949–50 (lasting from July 1, 1949 to June 31, 1950) amounted to $110 million. The dollar value of GNP for the same fiscal year, estimated by Nathan for the report, was $1,770 million. Net imports therefore were about 6 percent of the GNP for the fiscal year. Since the dollar value of aid imports in the calendar year 1950 was only $58.7 million (Bank of Korea 1956, p. 9) and that of the calendar year 1949 was $116.5 million, it hardly seems plausible that imports in 1949 were as large as 3.1 billion hwan.

[6] Despite his statement to this effect (Makinen 1971, p. 1237), Makinen now explains that that is not quite what he did. He apparently applied the official rate (9 hwan per dollar) to the commercial imports; but applied another conversion rate (25 hwan to a dollar) in evaluating the aid portion of imports. I find this ad hoc practice hard to accept.

All things considered, Makinen's error was in understating the 1950 GNP by about 40 percent and in overstating the net imports in 1949 by nearly 200 percent. Since his annual GNP figures, except for 1950, are generally consistent with the ones we obtained, we substitute the averages of two estimates of 1950 GNP given above (27.8 billion hwan) for the value of 16.3 billion hwan he had in table 14. Then we substitute the revised estimate for 1949 of imports of 1.2 billion hwan and exports of 0.1 billion hwan and those for 1950 of 1.5 billion hwan and 0.3 billion hwan respectively (ibid.). When the expenditure velocity is recomputed on the basis of these revised figures, its value rises from 100 of 1949 to 109 or 123 in 1950, depending upon whether the June and December averages or the quarterly averages of the money supply are used (Bank of Korea 1953, p. 8).[7] That it is more likely to have risen by something like 10–20 percent in 1950, rather than falling by 44 percent, is further supported by the fact that the Seoul consumer price index rose from 100 in June 1950 to 211 by July and to as much as 415 by September of the year (Bank of Korea 1964, p. 57). Since the money stock increased by no more than 22 percent between June and September (Bank of Korea 1953, p. 8), the real stock of money must have declined to about one-third of what it was before the war (i.e., June 1950).

At any rate, the general impression one gets is that, instead of letting actual balances run ahead of desired balances, the public appears to have promptly adjusted its cash balances following the outbreak of the war; and, for all we know, it continued to do so until the armistice was reached in 1953. There is no evidence that a long lag occurred in the adjustment of price expectations. If anything, the evidence points to a contrary conclusion.

References

Bank of Korea. *Monthly Statis. Rev.*, vol. 7, no. 1 (January 1953).
———. *Monthly Statis. Rev.*, vol. 10, no. 5 (May 1956).
———. *Monthly Statis. Rev.*, vol. 19, no. 3 (March 1965).
———. *Price Statistics Summary.* Seoul: Bank of Korea, 1964.
Campbell, Colin D., and Tullock, Gordon. "Some Little-understood Aspects of Korea's Monetary and Fiscal Systems." *A.E.R.* 48, no. 3 (June 1957): 336–49.
Friedman, Milton. "Prices, Income, and Monetary Changes in Three Wartime Periods." *A.E.R.* 42, no. 2 (May 1952): 612–25.
———, ed. *Studies in the Quantity Theory of Money.* Chicago: Univ. Chicago Press, 1956.

---

[7] For the purpose of economizing the space, a table containing various revised figures is deleted from the text. The table can be made available upon request.

Makinen, G. E. "Economic Stabilization in Wartime: A Comparative Case Study of Korea and Vietnam." *J.P.E.* 79, no. 6 (November/December 1971): 1216–43.

Nathan, Robert R., Associates. *An Economic Programme for Korean Reconstruction.* Seoul: United Nations Korean Reconstruction Agency, 1954.

The Real Stock of Money and Velocity in a Wartime Economy: A Reply

G. E. Makinen

Wayne State University

Professor Lee's comment is welcomed for it corrects an error in my original paper and shows that velocity did not fall in 1950 and, thus, my strong rationalization for its supposed behavior in terms of Cagan's two-lag hypothesis is incorrect. However, he is only partially correct in pinpointing the source of my error. At the time my estimates were originally made, I did not have access to or knowledge of the excellent facilities of the Asian Library of the University of Michigan. Subsequent research after receiving Lee's paper convinces me that the estimate he questions is in error. Therefore, a brief explanation as to the source and manipulation of my data is necessary to dispel the impression that they were "manufactured," and to pinpoint the error correctly.

Three values must be estimated to calculate what I term "expenditure velocity": the stock of money, nominal GNP, and the nominal value of exports and imports. While we used slightly different money series, I am unable to find the source of the difference. However, the valuation of nominal GNP and the foreign trade sector is where the errors are to be found.

The GNP data for 1949–53 were taken from H. B. Lee (1968, p. 57), which is incorrectly cited as *Bank of Korea, Economic Statistics Yearbook 1963.*[1] As his table contained internal errors, my corrected version appears as Makinen, table 12, (1971, p. 1235). Had I not made the corrections I would have obtained the same percentage changes in real GNP for 1949–50 (−15.1 percent) and for 1950–51 (−6.1 percent) as reported by J. K. Lee.

These constant hwan GNP data for 1949–55 were converted to current hwan by using the Seoul wholesale price index in the absence of an available GNP deflator. For the crucial year 1950, I used an average annual price index number which was obtained from H. B. Lee (1968, table 15, p. 59). Other data available confirmed that the price relatives

---

[1] The correct citation should have been *Monthly Statistical Review*, Bank of Korea (March 1965), pp. 19–20. Footnote 4 in J. K. Lee's paper is misleading, for the data provided in this citation make it quite easy to construct the GNP totals for the 1948–53 period.

[*Journal of Political Economy*, 1974, vol. 82, no. 6]

© 1974 by The University of Chicago. All rights reserved.

in that table for 1952–54 were the annual averages for the Seoul wholesale index. Subsequent research now shows that the relative for 1950 of 334 and for 1951 of 2,194 were, respectively, the averages for the first six months of 1950 and the last nine months of 1951. Their use thus resulted in a substantial undervaluation of nominal GNP for 1950 and a modest overvaluation for 1951.

As J. K. Lee and I both used the same dollar figure for imports and exports during 1949, the exchange rate is at issue. In the footnote to my controversial table 14, I said the official exchange rate was used. This is only partially correct. This was the rate used to convert imports and exports financed by the Korean government. However, for imports denoted as foreign economic aid and relief goods, I used the counterpart exchange rate which I assumed from incomplete information (Economic Planning Board 1962, p. 263) to be 25 hwan to the dollar when in fact it averaged 4.7 to the dollar in 1949 and 17 to the dollar in 1950.[2] My reason for using this so-called ad hoc method was to try to obtain a value for those imports which would measure their effect on velocity. The counterpart rate represented in essence the price paid by the Korean government for those goods. It should be noted that the official and counterpart rates were identical for 1952 through 1955 and fairly close in 1951. Only in 1949 and 1950 do they diverge noticeably. Nevertheless, my use of an incorrect value for that rate imparts a substantial upward bias to the import estimate for 1949 and a modest upward bias for 1950.

Thus, the sources of my incorrect velocity estimates are to be found in the use of an understated price index and an overstated exchange rate. However, it should be remembered that Lee's reconstruction of my velocity estimates does not in any way alter the conclusion of that section of my article: The rise in velocity in Korea contributed to the destabilization of the Korean economy whereas this was not true in South Vietnam. Rather, they reinforce that conclusion.

References

Economic Planning Board. *Korean Statistical Yearbook*, 1962. Seoul: Bur. Statis., 1962.

Lee, H. B. *Korea: Time, Change and Administration.* Hawaii: East-West Center, 1968.

Lee, J. K. "The Real Stock of Money and Velocity in a Wartime Economy." *J.P.E.* 82, no. 6 (November/December 1974): 1281–85.

Makinen, G. E. "Economic Stabilization in Wartime: A Comparative Case Study of Korea and Vietnam." *J.P.E.* 79, no. 6 (November/December 1971): 1216–43.

---

[2] The 1949 import value of 3.1 million hwan was thus obtained by converting the $116,500,000 in aid imports to 2,912,500,000 hwan (exchange rate 25 to 1) and adding to this sum the 113,009,000 hwan of imports financed by the Korean government.

Miscellany

What Is the Optimal Level of Pollution?

On a theoretical plane, the question has not been properly answered. Its neglect may be partly explained by its seeming obviousness. And though the answer is indeed obvious once it is correctly stated, it is not the answer one gleans from a reading of the relevant literature.

Other factors have probably contributed to its neglect. One is the altogether commendable concern now growing among economic specialists, particularly in the United States, about practical methods for stemming rising levels of air and water pollution. Another reason is that, in theoretical discussions much attention is given to what I have called (1965) "external diseconomies *internal* to the industry," or to the "activity," the phenomenon associated chiefly with the names of Allen Young and Frank Knight (1924), whose contributions are summarized in the classic Ellis and Fellner paper (1943). The "external diseconomy" here springs from nothing less familiar than the eventually diminishing returns to a variable factor arising from the fixity of other factors. Where—as in the case of agriculture—the scarce factor, land, is assumed properly priced—that is, by reference to the difference between the average and marginal product of labor—the resulting competitive supply curve is also the correct marginal cost curve of grain, and the perfectly competitive output is therefore the optimal output. Where, instead, the scarce factor is some common property such as a road or a fishing area, it is not priced at all. The competitive supply curve of the good produced on, or with, this common property factor is only an average curve, one that excludes rent to the scarce factor. Consequently, the competitive equilibrium (of traffic or output) exceeds optimality.

Restriction of one's attention to these external diseconomies that are internal to the industry, and therefore also to the theme of an unpriced scarce factor, suggests that the optimum external diseconomy to be borne by the industry (and by society) is that associated with the optimal output

[*Journal of Political Economy*, 1974, vol. 82, no. 6]
© 1974 by The University of Chicago. All rights reserved.

of the corrected supply curve—one that is an average curve that includes rent to the scarce factor. For all these "internal-to-the-industry" cases the answer is satisfactory simply because it is assumed that the external diseconomy can be varied only by varying the industry's output. Such an assumption, however, is not justified for the class of external diseconomies that are external to the industry or activity, especially those which, under the umbrella term "pollution," have engrossed the public's attention over the past decade. Though, again, an unpriced common property resource, such as air or water, is at the base of the problem, there are now ways of varying the amount of the externality other than by varying the output of the good itself. It is this fact—that a pollutant externality can be reduced in a number of ways—that calls for a more general statement about the optimum level of pollution.[1]

---

[1] If ever an account of the development of this aspect of economic thought comes to be written, there is sure to be some speculation about the long neglect of a general statement about the optimum level of a pollutant externality. My guess is that one reason for it has been the intellectual energy absorbed by the reaction against the Pigovian doctrine, and especially against the classic tax-subsidy solution. In this connection one thinks immediately of the famous Coase monograph (1960), and also perhaps of the short piece by Buchanan and Stubblebine (1962), which together seemed to have started an iconoclastic movement to edge the master from his niche in the hall of fame. Although there have been earlier criticisms of parts of Coase's analysis (Mishan 1965) and also of Buchanan and Stubblebine's paper (Shibata 1972), the Pigovian doctrine and the tax-subsidy solution have been vindicated in an elegant and incisive article by Baumol (1972), which gently but firmly took these writers to task for discovering "errors" in the Pigovian doctrine that were the result wholly of a lack of care in their own analyses. As for other critics of the classical doctrine who reached conclusions seemingly at variance with those of Pigou, the novelty of their findings has arisen simply from the novelty of their assumptions. To mention those that spring to mind, Plott (1966) concocts a counter-Pigovian (and counterintuitive) example of optimality, one requiring an expansion of the polluting activity, by assuming a production function that uses less of the factor directly responsible for the pollution as more output is produced. Dolbear (1967) uses a model in which changes in welfare affect the valuation of, and demand for, goods and bads, an assumption that—not surprisingly—undermines the uniqueness of an optimal solution (see also Mishan 1967). Schall (1971) manages to produce an optimal increase in the market output of a good that generates an external diseconomy by assuming a simultaneous contravention of both factor efficiency and top-level efficiency conditions. For it is then conceivable that a correction of both conditions entails so large an improvement of productive efficiency that the resulting increase in output of the good in question swamps the movement that would otherwise reduce the output of this good originally having a marginal social cost greater than its price. In view of Mishan (1965) and Schall (1971), at least, the statement by Goetz and Buchanan (1971)—namely, that "the impact of production externalites has heretofore been conceived of exclusively in terms of a sub-optimal exchange equilibrium on the production possibility frontier"—is invalid. Not only do they fail to observe that the simultaneous infringement of productive and top-level efficiency conditions occurs when external diseconomies are internal to the industry, but they go to the unnecessary trouble of contriving a model in which—for some unexplained reason—the unit cost of the ith firm depends not on the output of all firms, but only on the output of the remaining firms. (This implies a cost advantage in producing a total output by a number of firms which is smaller than that in the uncorrected market equilibrium.)

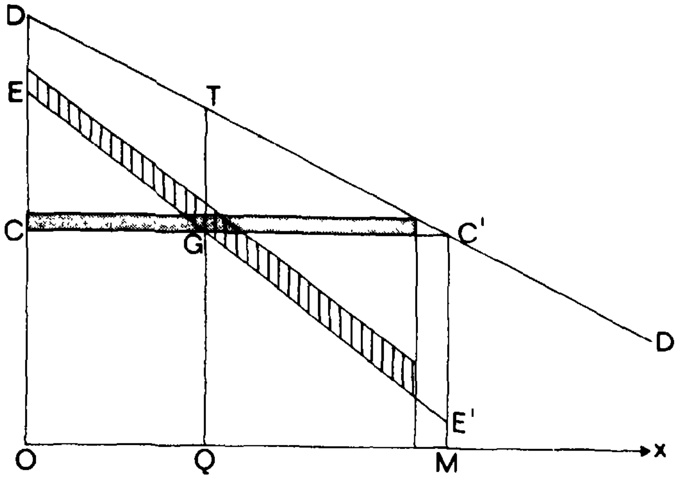


FIG. 1

I

Though not necessary to the conclusions, in the interests of simplicity the analysis which follows is strictly partial: in all other sectors outputs are already optimal, and our concern is then solely with reaching an optimal solution in this one industry or activity that produces an external diseconomy. To avoid obvious modifications, we shall assume that the welfare effects on the amounts demanded are zero; initially that there are no transactions costs broadly defined; and that the industry which produces an output X is perfectly competitive and has a constant long-period supply curve. If the reader can bear the tedium, we shall suppose that the external diseconomy produced in conjunction with X is, again, smoke, and that the value of the social damage associated with any increment of output is measured as the vertical distance between the demand curve DD' and the curve EE' in figure 1.

The conventional analysis goes as follows. With the demand curve DD', the uncorrected long-period market equilibrium output of X is equal to OM. The "optimal output" OQ occurs where the resultant marginal social valuation measured by curve EE' is equal to the marginal resource cost CC', an output that can be realized by imposing an "optimal tax" TG on each unit of X.[2] The "optimal pollution," which has to remain in the economy, and be suffered by the pollution victims, is that associated with output *OQ*, and valued, or rather "costed," by society at a sum equal to the area *EDTG*.

---

[2] It is yet more conventional, in the Anglo-Saxon literature, to add the value of the smoke damage to the cost curve so as to construct a marginal social cost curve rather than, as here, to subtract the value of smoke damage from the market valuation of the goods produced in association with it. But the analysis which follows is simpler to follow by adopting the latter construction.

These conclusions would be satisfactory if there were no methods of dealing with the smoke other than by varying output of *X*, although the *TG* excise tax would be only one device through which this "optimal" output *OQ* could be determined. For, in the last resort, quantitative controls alone could be used to ensure that no more than *OQ* of *X* is produced. Moreover, bargaining between the industry and the smoke victims would also result in *OQ* of *X* being produced, irrespective of which party under the law had to pay the other to take action.

But, as mentioned above, there can be a number of ways of curbing smoke pollution, and Coase's (1960) article has at least the merit of drawing our attention to this mundane fact. In section 9 of the article, for instance, he assumes a case in which the damage caused by the smoke is $100 per annum, but in which there is a smoke-preventing device costing $90 and also the option of moving all the smoke victims from the locality for the modest sum of $40 per annum.[3] For that matter, the relocation of the smoky factory to a more sparsely inhabited area may also be feasible. Clearly, there is no limit to the number of possible ways of reducing the pollution, but nothing is lost in generality of conclusions by confining ourselves to three forms of reducing it: A, the one most familiar to economics students, a reduction of the output of *X*; B, the employment of any particular smoke-preventive technique; and C, the movement elsewhere of the smoky industry.

II

The impression conveyed by the literature on externalities that occasionally recognizes alternative methods of dealing with them—certainly in the Coase article but also in those of others[4]—is that where one method is not clearly more economic than the others, practical circumstances will suggest which is to be adopted. Thus, in the Coase example referred to, the most economical solution is that of moving the smoke sufferers to a smokeless area, whereas, when it comes to practical considerations, the economic specialists have a marked predilection for an effluent tax, at least for water pollution.[5] In contrast, the theoretical solution to the problem of the optimal amount of pollution implies that if there are *n* methods of varying the pollution, then, in general, all *n* methods are to be simultaneously employed.[6]

---

[3] As Baumol (1972) pertinently observes, however, if it costs the victims but $40 per annum to move elsewhere, they would of their own volition do so rather than put up with smoke damage that they reckon as equivalent to $100 per annum. Once they move away, no further smoke damage is suffered by people in the vicinity, and the question of a tax on the output of the smoky factory no longer arises.

[4] This incidental impression is clearly conveyed by a splendid summing up of this aspect of the literature by Davis and Kamien (1969, pp. 67–86), and also in Baumol's paper (1972, pp. 313–14), where, in his discussion of "multiplicity of local optima," he suggests that we choose the maximum among them. The same impression arises in reading the otherwise excellent article by Baumol and Bradford (1972). Thus, toward the end of their section 4 on "Spatial Separation as a Palliative," they state: "These diagrams illustrate the proposition that sufficiently severe externalities make locational specialization economical."

In elaborating this conclusion, there are advantages in proceeding as follows. First, the "optimal" B and C methods will each be summarized before being compared with the A solution. Following this, each method is compared with what we may call the "composite" optimum solution inasmuch as this solution employs each of the three methods simultaneously, though using each method less intensively than if that method were used alone and to the exclusion of the other methods. Our results are then formalized, and some of their policy implications are briefly considered.

While the A method is the one which reduces the total increment of pollution that is generated by a unit of output of *X* by the expedient of reducing the production of that unit of *X*, the B method involves expenditure on resources in creating devices that act to reduce, in some degree, the existing level of pollution on all the *X* being currently produced. Thus, an initial increment of expenditure on some preventive technique will be worthwhile if it reduces an amount of pollution (generated by the existing output of *X*) that confers a social gain greater in value than the increment of expenditure. A marginal reduction in the value of social damage arising from a marginal reduction in the pollution generated by output *OM*—in effect the marginal social gain arising from a marginal decrease in overall pollution—can be represented in our figure 1 by the area of the shaded strip lying on top of *EE'*. The (smaller) cost of effecting this social gain can be an overhead expense to each firm in the short period, but in the long period the industry supply curve is raised by the dotted strip lying on top of *CC'*. Since this adjustment raises the equilibrium price of *X*, the equilibrium output of *X* will be somewhat smaller than *OM* as indicated in the figure. Clearly, one may continue to improve matters, that is, to make net social gains, by incurring successive increments of pollution-abatement expenditure until the cost of doing so is no longer exceeded by the corresponding "strip" of social gain.

The effect of the "optimal" solution by method B could then be

---

[5] The reasons for this are nicely summarized in the article by Freeman and Haveman (1971).

[6] I say "in general" since the use of one method may, for technical or other reasons, preclude the use of one or more other methods. Also, some methods may be too costly to adopt for the particular levels of pollution in question. Thus, for all *n* methods to be simultaneously adopted, certain conditions would have to be met (which we can assume are met in our three-method model) in order for each to be economically feasible.

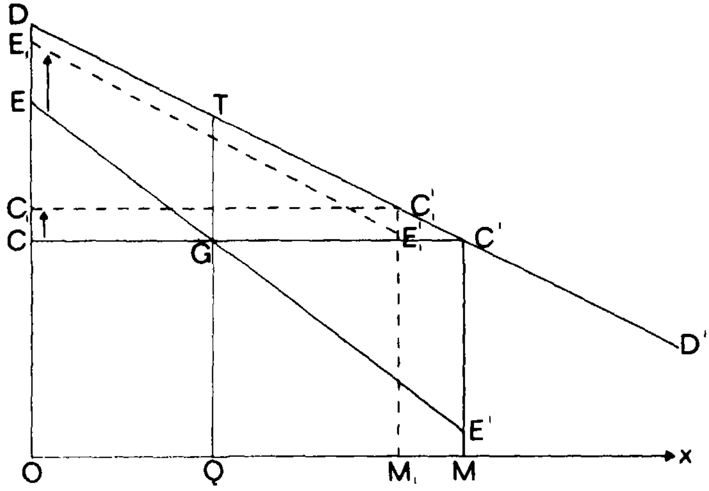


FIG. 2

represented on the same diagram in figure 2. Total expenditure on pollution abatement has raised long-period marginal cost to C_1C_1', thereby reducing equilibrium output of X to OM_1. The value of the pollution remaining—the "optimal" pollution, since the marginal gain of pollution reduction by method B is equal to the marginal cost so incurred—is given by the area $DE_1E_1'C_1'$, and this may be greater or less than $EDTG$, the "optimal" pollution that remains when using the A method alone.

Method B, involving preventive techniques, is implied by proposals for pollution or effluent charges, or by taxes designed to reduce pollution to some arbitrary "tolerable" level.[7] Obviously, bargaining as between the industry and the victims of pollution—about the overall reduction of pollution to be aimed at by preventive techniques—will also result in a B-method "optimal" level of pollution.

III

The diagram by which the "optimal" pollution on the B method is commonly exhibited is that shown as figure 3. Along the horizontal axis, units of pollution are measured, and along the vertical axis is measured the marginal social value of the pollution damage and also the marginal

---

[7] Though such a tax implies that firms coping with pollution with differing degrees of efficiency will undertake to reduce different amounts of pollution—so that their combined costs for reducing a given amount of pollution are minimal—this optimal aspect of such a tax does not emerge in our simple model in which all firms are assumed to be of equal efficiency.

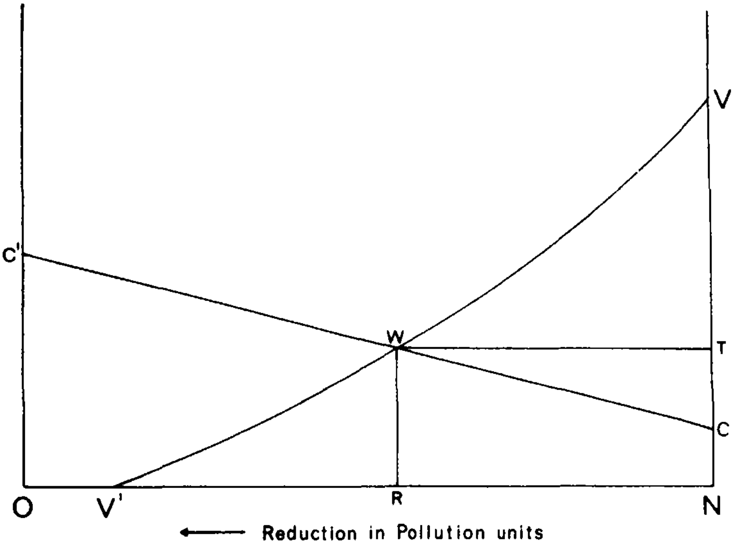


FIG. 3

opportunity cost of pollution abatement.[8] Thus, *ON* is the total pollution initially generated by the uncorrected *OM* market output of *X*, and *OV'V* is the marginal social damage curve. Up to *OV'* units the pollution has no perceptible impact on people's welfare, after which successive increments of pollution entail increasing social loss—measured, say, by the maximum sum society is willing to pay to avoid that unit of pollution—the *n*th unit of pollution suffered being "valued" at *VN*. If we look at the diagram instead from right to left, the *VV'O* curve can be construed as the downward-sloping valuation or demand curve for pollution reduction, and *CC'* as the upward-sloping marginal cost curve of pollution reduction. The area *VNV'* in the figure is equal to area *DEE'C'* in figure 2, and both areas are equal to the total value of social damage when nothing at all is done about the pollution.

Figure 3 clearly suggests that net social gain from reducing pollution is maximized by a reduction of *NR* pollution units, at which point the (falling) marginal gain from pollution reduction is equal to the (rising) marginal cost of pollution reduction. A tax of *TN* per unit of pollution would therefore be "optimal" since—without knowledge of, or reference to, the *VV'* curve—the levying of such a tax will impel the industry to reduce pollution by *NR*. For up to the first *NR* units of pollution, it is

---

[8] These marginal curves are sometimes derived, respectively, from anterior curves of total damage value of pollution and total cost of reducing pollution, as, for instance, in Ridker (1967).

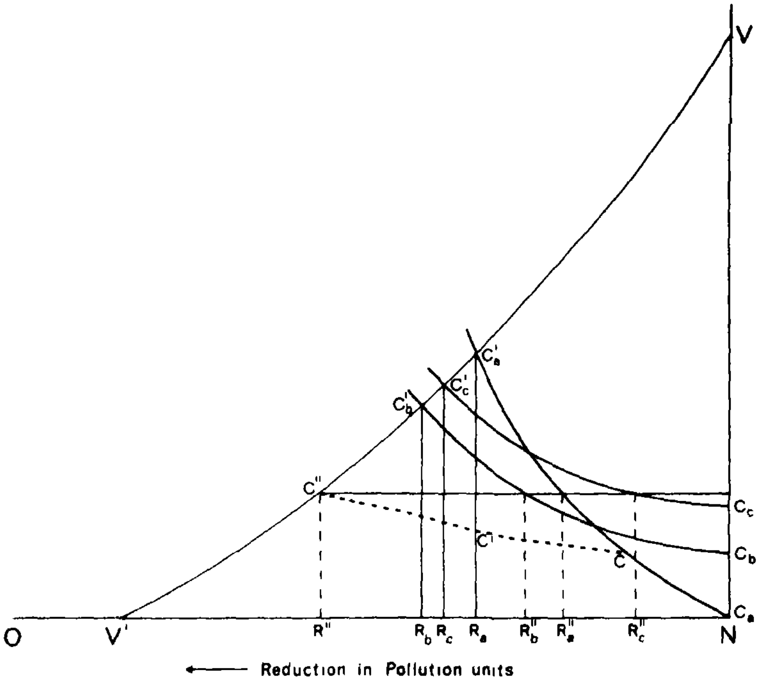


FIG. 4

cheaper to reduce pollution than to pay the tax. After *NR* units, the reverse is true. The "optimal" pollution remaining is, therefore, *OR*, and this is "valued" at a sum that corresponds to the area of triangle *V'RW*—the sum represented in figure 2 as equal to the area $DE_1E_1'C_1'$.

Now, by measuring along *NO* the distance in miles that the polluting *X* industry can be moved from the boundary of the inhabited area,[9] we can interpret the *CC'* and *VV'* curves, respectively, as the marginal cost per mile of movement and the marginal social gain per mile of that movement arising from the associated reduction in pollution. We should then conclude that *NR* is the "optimal" distance to move the industry. However, since it will be convenient to represent this method C on the same axes as method B, we assume the possibility of a unique transformation of distance into units of pollution reduction. Thus, in figure 4, the horizontal axes measures units of pollution, and the marginal cost curve for method C with respect to these units appears as C_cC_c', which curve can now be directly compared with the marginal cost curve C_bC_b' for method B, which, like the marginal valuation curve VV', is the same as that in figure 3.

---

[9] Alternatively, the horizontal distance from *N* can measure the number of factories to be moved some fixed distance away. It is part of our general conclusion that a combination of both methods can produce a better result than either alone. Although movement in some geographical directions will be more efficient than in others in reducing the effect of pollution on the inhabited area, there is nothing to be gained at this stage in making the analysis more "real" in these respects.

A similar marginal cost curve for method A can be shown on the same diagram as C_aC_a'. This is in effect a marginal opportunity curve derived primarily from the foregoing of excess benefit—demand price less marginal resource cost—as the output of X is reduced from OM toward zero.[10] Transforming reduced output of X into reduced pollution, simplified by assuming a fixed proportion between smoke and output, produces this required cost curve of reducing pollution by method A with respect to pollution units.

In general, the "optimal" pollution levels when each one of the three methods is used alone will differ. In figure 4 they are OR_a, OR_b, and OR_c, respectively, for methods A, B, and C. Their corresponding "optimal" pollution taxes also differ.[11] But if all three methods of reducing pollution are to be simultaneously employed, then economy requires that the amount reduced by each method is such that the marginal cost of each method is the same. This result follows from the idea of reducing the first, second, third, . . . *n*th unit of pollution by whichever method is the cheapest for that unit, bearing in mind that eventually each of the methods is increasingly used. The succession of lowest incremental costs so derived forms the composite marginal cost curve, and the true optimal pollution reduction is determined where this curve cuts the VV' curve at R''.

For simplicity of construction, assume the marginal cost curve of each of our three methods is independent of the costs, if any, incurred by the other two methods.[12] We can then construct the composite marginal cost curve for pollution reduction, $C_aCC'C''$, by "horizontally adding together" the three separate marginal cost curves. From C_a to C on this composite curve, method A only is used to reduce pollution. From C to C', some reduction is effected by both A and B methods. From C' onward, further reductions receive a contribution from each of the three methods.

The reader will at once observe that the composite optimal pollution that remains, $V'R''$, is smaller than the "optimal" pollution determined by the employment of any one method alone. Consequently, the composite pollution tax $C''R''$ is also smaller than a pollution tax determined by reference to any single method of pollution reduction. Since the

---

[10] In perfect competition, equilibrium output is M where long-run marginal cost is equal to price. At this output M, the marginal opportunity cost of reducing the output of X is therefore zero.

[11] The "optimal" pollution tax for the A method alone, equal to $C_a'R_a$ in fig. 4, can be transformed into an excise tax on good X, and the "optimal" pollution tax for the C method alone, equal to $C_c'R_c$ in fig. 4, can also be transformed into a zoning tax based on mileage from the inhabited area.

[12] The assumption facilitates the geometry, but, as we shall see later, the resulting marginal conditions remain unaffected when we assume, more realistically, that the marginal cost curve of any one method is affected—presumably shifted upward—by an increase in the use of the other methods (see n. 15).

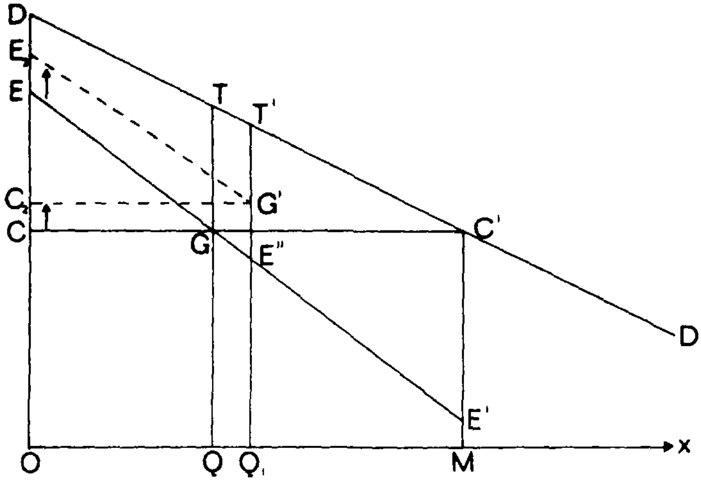


FIG. 5

magnitude of the composite pollution tax is to be adopted by each of the three methods when they are used together, the pollution reduction contributed by each of the methods when they are used together is smaller than the "optimal" pollution reduction effected by that method when it is used alone. Thus, of the total pollution reduced jointly, NR'', methods A, B, and C, are responsible, respectively, for amounts NR''_a, NR''_b, and NR''_c. Clearly, each of these amounts is smaller than NR_a, NR_b, and NR_c, these being, respectively, the "optimal" pollution reductions of methods A, B, and C, when each method is used alone.

It may be revealing to depict the effects of this composite pollution optimum also on the original demand-and-supply diagram. The composite pollution tax implies an excise tax on the output of X that is smaller than TG in figure 2.[13] It is equal, say, to $T'G'$ in figure 5 (derived by reference to $C''R''$ in fig. 4). The optimal output of X resulting from applying the composite pollution tax to the A method is, therefore, OQ_1, which is larger than the output OQ corresponding to the use of method A alone. The optimal employment of methods B and C has, between them, raised the unit cost of this output from OC to OC_2 and has reduced pollution damage by a value equal to the area $EE_2G'E''$. The value of the optimal pollution remaining, the area $DT'G'E_2$, is then also equal in area to triangle $V'R''C''$ in figure 4.

These findings are easily formalized and extended. At any moment of time there will be a number of economically feasible technologies that can be used directly to reduce the level of pollution generated in the production of a given output of good X, including perhaps one or more methods of recycling the pollutant,[14] each having its own marginal cost curve of pollution reduction. Although under suitable conditions all these direct methods of pollution abatement will be simultaneously used, there can be indivisibilities of capital expenditure that demand an initial choice among the alternatives, one based on a comparison of estimates of the total costs to be incurred in reducing some expected amount of pollution. But for all these direct methods of pollution abatement that are simultaneously employed, the composite optimal solution requires that the marginal social gains from reducing pollution be equal to the marginal composite cost of reducing pollution, and equal also to the marginal cost of reducing pollution by each of the methods in use.[15]

---

[13] The TG can also be derived from pollution tax C'_aR_a in fig. 4.

IV

In a world in which transactions costs, broadly defined, are low in relation to net social gains from reducing pollution, an ideal institution, on grounds both of equity and allocation, would be a law that placed on all polluting manufacturers the obligation to compensate all pollution victims. Businesses would then have a direct incentive to seek new and more economical ways of reducing pollution damage in order to reduce their total costs, including residual compensatory payments. Since the conflict today is between polluting industry on the one hand and the

---

[14] Recycling can be regarded as a combination of pollution reduction and goods production. The marginal social value of the goods that emerge in the recycling process is to be subtracted from the appropriate marginal cost of the particular recycling process to yield the relevant marginal cost of reducing pollution by this means.

[15] Let Z_1, Z_2, Z_3 be reductions in amounts of pollution, and total cost

$$Ci = Ci(Z_1, Z_2, Z_3). \qquad i = 1, 2, 3.$$

Assuming that $\partial C_1/\partial Z_1, \partial C_2/\partial Z_2, \partial C_3/\partial Z_3 > 0$, it can be shown that unless $\partial C_1/\partial Z_1 = \partial C_2/\partial Z_2 = \partial C_3/\partial Z_3$, then $C(Z) = C_1 + C_2 + C_3$ is not a minimum. Thus, if for $\bar{C}(Z) = \bar{C}_1 + \bar{C}_2 + \bar{C}_3$, where $Z = \bar{Z}_1 + \bar{Z}_2 + \bar{Z}_3$, we have $\partial \bar{C}_1/\partial \bar{Z}_1 > \partial \bar{C}_2/\partial \bar{Z}_2 = \partial \bar{C}_3/\partial \bar{Z}_3$, we can find different values for Z_1, Z_2, and Z_3 which also add up to Z but which yield a $C(Z) < \bar{C}(Z)$. Let $dZ = dZ_1 = dZ_2$; then $Z = (\bar{Z}_1 - dZ_1) + (\bar{Z}_2 + dZ_2) + \bar{Z}_3$ also. For this combination of Z_1, Z_2, and Z_3 we have

$$C(Z)\left[= \left(C_1 - \frac{\partial \bar{C}_1}{\partial \bar{Z}_1} dZ_1\right) + \left(C_2 + \frac{\partial \bar{C}_2}{\partial \bar{Z}_2} dZ_2\right) + C_3\right] < \bar{C}(Z)\,[= \bar{C}_1 + \bar{C}_2 + \bar{C}_3],$$

that is,

$$\left[C_1 + C_2 + C_3 + \left(\frac{\partial \bar{C}_2}{\partial \bar{Z}_2} - \frac{\partial \bar{C}_1}{\partial \bar{Z}_1}\right) dZ\right] < [\bar{C}_1 + \bar{C}_2 + \bar{C}_3],$$

since, by assumption, $\partial C_2/\partial \bar{Z}_2 - \partial \bar{C}/\partial \bar{Z}_1 < 0$. We limit ourselves here to the necessary conditions only. Thus, we ignore the possibility that the overheads associated with any one method may exceed the net social gains from its employment; also, the possibility that the difference in overheads as between two methods may be so great that it may be economical to withdraw entirely the one with the larger overheads.

public at large on the other,[16] the transactions costs associated with compensating the victims under such a law are likely to be heavy, including, as they must, the costs of information, of identifying claimants, of legal fees, of negotiating fees, etc. Unless the available methods are so efficacious as to leave a negligible amount of pollution, such a law might cause a large number of polluting enterprises to close down altogether. And this "polar" solution, though economically justifiable under such a law, may be economically worse, or better, than the opposite polar solution, that of no reduction at all of pollution, which comes about if the effect of existing laws is to place the burden of compensation wholly on the public at large.[17]

Better than either of these polar solutions are likely to be those that result when governments undertake some initiative and, therefore, come to bear a part of the consequently smaller magnitude of transactions costs (interpreted to include research, monitoring, and otherwise enforcing the law) that are involved in nothing more ambitious than maintaining "tolerable" standards. As some economists have observed,[18] the costs associated with gathering the data necessary to estimate the $V'V$ curve in order to calculate the ideal tax could be disproportionately high compared with the anticipated value of the net social gains. It is then more practical to devise means of reaching agreement on standards—say the level of maximum "tolerable" pollution—and calculating or discovering (by trial and error) taxes that give effect to them.

The foregoing analysis suggests that we bear in mind also that enterprising firms will, given the tax or other incentive, continue to seek ways of transforming the pollution tax into an output tax, into a zoning tax,[19] and so on, and that they will find it most economical to operate on all these "antipollution fronts" simultaneously. We might then remind ourselves that for any required reduction in pollution, the tax set by reference to the simultaneous use of a variety of methods will be lower than that set by reference to any single method of pollution reduction.

E. J. MISHAN

London School of Economics

---

[16] Since costs of pollution reduction will, in any case, be passed onto the purchasing public by the manufacturers, it is more accurate to talk of a conflict between people as consumers of the product and people as the inhabitants of the environment. Nonetheless, manufacturers can be regarded as representatives, in the first instance, of the interests of people as consumers.

[17] See Mishan (1971, sec. 10).

[18] For a recent paper favoring the adoption of pollution standards, see Baumol and Oates (1971).

[19] For the pollution tax effectively to become a zoning tax also, it is necessary that the tax on a particular pollutant be specified as that pollutant falling within a given inhabitable area.

References

Baumol, W. J. "On Taxation and the Control of Externalities." *A.E.R.* 62, no. 3 (June 1972): 307–22.

Baumol, W. J., and Bradford, D. F. "Detrimental Externalities and Non-convexity of the Production Set." *Economica* (May 1972), pp. 160–76.

Baumol, W. J., and Oates, W. E. "The Use of Standards and Prices for Protection of the Environment." *Swedish J. Econ.* (March 1971), pp. 42–54.

Buchanan, J. M., and Stubblebine, W. C. "Externality." *Economica* 29 (November 1962): 371–84.

Coase, R. H. "The Problem of Social Costs." *J. Law and Econ.* 3 (October 1960): 1–44.

Davis, O. A., and Kamien, M. I. "Externalities, Information, and Alternative Collective Actions." In *The Analysis of Public Expenditures: The PPB System.* Washington: Government Printing Office, 1969.

Dolbear, F. T., Jr. "On the Theory of Optimal Externality." *A.E.R.* 57, no. 1 (March 1967): 90–103.

Ellis, H. S., and Fellner, W. "External Economies and Diseconomies." *A.E.R.* 33 (September 1943): 493–511.

Freeman, A. M., and Haveman, R. H. "Clean Rhetoric, Dirty Water." *Public Interest* 28 (Summer 1972): 51–65.

Goetz, C. J., and Buchanan, J. "External Diseconomies in Competitive Supply." *A.E.R.* 61 (December 1971): 883–90.

Knight, F. H. "Some Fallacies in the Interpretation of Social Cost." *Q.J.E.* 37 (August 1924): 582–606.

Mishan, E. J. "Reflections on Recent Developments in the Concept of External Effects." *Canadian J. Polit. Econ.* 31 (February 1965): 3–34.

———. "Pareto Optimality and the Law." *Oxford Econ. Papers* 9, no. 3 (November 1967): 255–87.

———. "The Postwar Literature on Externalities: An Interpretative Essay." *J. Econ. Literature* 9 (March 1971): 1–28.

Pigou, A. C. *The Economics of Welfare.* 4th ed. London: Macmillan, 1946.

Plott, C. R. "Externalities and Corrective Taxes." *Economica* 33, no. 129 (February 1966): 84–87.

Ridker, R. G. *The Economic Costs of Air Pollution.* New York: Praeger, 1967.

Schall, L. D. "Technological Externalities and Resource Allocation." *J.P.E.* 79, no. 5 (September 1971): 983–1001.

Shibata, H. "Pareto-Optimality, Trade, and the Pigovian Tax." *Economica* 154 (May 1972): 190–202.

Poor Old Capitalism: A Review Article

Evsey D. Domar
Massachusetts Institute of Technology

I

For a teacher of comparative economic systems, the publication of *The Capitalist System: A Radical Analysis of American Capitalism*, by Edwards, Reich, and Weisskopf, is a windfall: it presents him with 540 pages of radical attack on capitalism all in one place, including passages from Marx and Engels, Polanyi, Dobb, Baran and Sweezy, Mumford, Gintis, Fromm, Bowles, and others.[1] The three editors wrote brief introductions to each chapter and to each selection and contributed several articles of their own. Most of the material is nontechnical and well written; it should be accessible to any intelligent undergraduate and lay reader.

An ordinary book of readings is essentially a pedagogical tool. Its editor need not agree with, or be held responsible for, the views and conclusions presented in it. But this is no ordinary book of readings. It presents almost exclusively the radical point of view. The introductions written by the editors gave them ample opportunity to state their disagreements, if any, with the contributors.[2] Hence, the reviewer has the right to treat the book as an integrated work and to hold the editors responsible for its content.

Nevertheless, if the book's subtitle were *A List of Evils of the American Economy*, or something like that, I would not quarrel with it; I might even suggest a few extra ones. But what provoked me was the word "Analysis" in the subtitle. Now, I thought, our radicals will analyze the causes of American evils and show that they are indeed produced by capitalism. Unfortunately, my high hopes were disappointed.

---

[1] Richard C. Edwards, Michael Reich, and Thomas B. Weisskopf, *The Capitalist System: A Radical Analysis of American Society* (Englewood Cliffs, N.J.: Prentice-Hall, 1972).

[2] Their disagreement with a passage from Baran and Sweezy was duly recorded on p. 309.

[*Journal of Political Economy*, 1974, vol. 82, no. 6]
© 1974 by The University of Chicago. All rights reserved.

II

American capitalism is charged with six principal evils: inequality, alienation, racism, sexism, irrationality, and imperialism, plus a number of lesser ones. But before discussing them, let me make a few general comments.

An analysis of capitalism, like any analysis, can be expected to consist of two parts: first, a logical formulation of a hypothesis showing how this or that evil is caused by capitalism; and, second, an empirical testing of the hypothesis against the reality of capitalist and noncapitalist systems.[3] There is no shortage of logical formulations, of different degrees of plausibility, in the book. But there is almost a complete absence of empirical verification.[4] Since the evils are both complex and not easily quantifiable, the authors (that is, the editors and the contributors) could not be required, at least at this stage, to come forth with a battery of regression equations, but surely, as a first step, they could have made an attempt to examine historical trends and to establish the presence or absence of each evil in other capitalist and noncapitalist countries. In particular—most fortunate for this attempt—there are now several socialist countries, some of them quite advanced and most of them sharing our common cultural background. On one of them—the Soviet Union—there exists a large literature in English, while the others have not been neglected either.

But no comparisons of any importance are made in the book. We discover that there is not a single socialist country in the world! The Soviet Union and the other East European countries are referred to as "state socialism" (pp. 4, 277, 281, 362, 524–25) or as "so-called socialist" (p. 277). They are treated with disdain and together with the state-capitalist countries (England, France, or Sweden) are declared not to be "model societies of socialism to be emulated" (p. 4). Worse than that, "The state socialist countries of the Soviet Union and Eastern Europe are to true socialism what 'the monsters of the paleolithic era are to present animal species: clumsy, abortive, prototypes'" (p. 4).

If countries which have been regarded by themselves and by others to be socialist have turned out to be something else, surely an explanation is in order. Since none is provided in the book, let me suggest two alternatives. (1) Lenin, Stalin, and, by implication, all other socialist leaders—including Tito and Mao—never intended to build socialism. (2) They did try, but failed miserably, ending up with "paleolithic monsters," in fact.

---

[3] It seems that there is no disagreement between the authors and myself on methodology, as the following quotation taken by them from a paper by Baran and Hobsbawm testifies: "It suggests the necessity of an interpretation of theory and concrete observation, of empirical research illuminated by rational theory, of theoretical work which draws its life blood from historical study" (p. 55).

[4] The three papers by Reich and Weisskopf, respectively, discussed below in Section III, do contain, or are based on, empirical data. Some data are also given in the interesting paper "The Negro Worker in the Chicago Labor Market" by Harold Baron and Bennet Hymer, pp. 297–305. There are also scattered statistical data elsewhere in the book.

I will leave the choice and the consideration of the sad implications of each alternative to the reader.[5]

Terminological arguments are, of course, fruitless, and I would not have started this one if our authors were consistent. But they are not. Thus, we read in the book about "a powerful socialist sector of the world" (p. 55) and about "opposition from the socialist world" (p. 409). There are "successful socialist revolutions" (p. 418), "development of socialism" (p. 418), and "rise of socialism" (p. 424), and there is the "military strength of the socialist camp" (p. 425). Of course, the several authors need not agree among themselves, but the failure of the editors to clarify this rather important issue in their introductory comments leaves the reader with the impression that socialism does and does not exist at the same time and place, depending on the need of the argument.

But whatever these countries are, they are not capitalist. Our authors' failure to investigate whether capitalist evils exist there as well places every statement in the book under suspicion of being a half-truth at best, like a savings bank's claim that it pays the highest rate of interest allowed by law. This difficulty is not removed by the editors' statement that "it is not our intention to imply that *all* forms of oppression are a result of capitalist institutions" (p. 5), and by their assertion that "the elimination of basic capitalist institutions is necessary, though not sufficient, to eliminate the oppressive problems of the modern world" (p. 5). At best, such face-saving afterthoughts leave the question wide open. At worst, they are grossly misleading.

III

The six evils cover so much ground that only a few comments on each can be made in my limited space. I will try to avoid factual disputes—there is not much serious factual information in the book in any case—and will concentrate instead on the analysis of the arguments presented and on comparisons, whenever possible, with socialist countries and particularly with the Soviet Union.[6]

Alienation.—In simple English, this word means that workers are dissatisfied because they neither own nor control the means of production, do not participate in decisions, and are compelled to perform repetitive and meaningless-to-themselves tasks. Since this dissatisfaction is a state of mind, it is not easily quantifiable, and sources differ on its importance and extent.[7] Our authors are sure that it affects very large numbers of workers. If so, it is rather strange that labor unions have shown so little interest in mitigating it. Perhaps they find that dissatisfied workers make better union members. In any case, one would expect that a book of this kind would devote at least one long chapter to the analysis and criticism of union policies. But, except for a few scattered remarks, little criticism of unions and even of the racial policies practiced by some of them is found in the book.[8]

---

[5] If existing socialist countries are not really socialist, could not the proponents of "real" capitalism declare all existing capitalist countries not capitalist? After all, none of them has perfect competition and complete laissez faire. It would be particularly pleasant to exclude Franco's Spain and the colonels' (or was it the generals'?) Greece, among others. Then we would be comparing completely imaginary systems!

[6] I concentrate on the Soviet Union because of its large size and the relative abundance of Western studies on its economy. Also, I am more familiar with its economic system than with those of other socialist countries.

With the exception of Yugoslavia, workers do not own and do not control the means of production in the socialist countries, either. With similar technology, they are also engaged in repetitive and uninteresting work. Socialist enterprises are also organized in a hierarchial manner, with the managers (directors) wielding great power. That some dissatisfaction exists among Soviet workers as well is confirmed by the recent publication of at least two serious Russian books on the subject (Osipov 1966; Zdravomyslov, Rozhin, and Iadov 1970).

Nevertheless, I would expect Soviet workers to be less alienated than ours, basically because of the chronic existence of excess demand. The producers (including workers) are less concerned with pleasing their customers by the proper assortment and quality of goods. They are also less concerned with profits. Hence, there is less pressure on the workers and greater job security, reinforced with usage and legislation. From his own experience with shortages, the worker may derive a feeling of accomplishment from increasing his output without endangering his job. The enterprise's welfare and cultural activities, particularly important in an economy of scarcities, create an additional bond. And of course he is told time and again that profits, far from enriching capitalist exploiters, are put to social use.

---

[7] According to the Survey Research Center of the University of Michigan, as cited by Jencks et al. (1972), the majority of workers seem to be quite satisfied. Thus, 63 percent would recommend their job to a friend; only 27 percent would hesitate to take the same job, and 9 percent would not take it; as many as 49 percent would choose the present job if they could have any job they wanted. A Gallup poll reported in the *New York Times* (December 6, 1973, p. 24) says that 79 percent of workers were satisfied with their jobs. This, however, was a decline from 85 percent recorded in 1963. On the other hand, a government report, *Work in America* (1972), gives the impression that the dissatisfaction among workers is much greater. Unfortunately, the data are presented in this report so sloppily that it is difficult to judge whether dissatisfaction affects many or few workers and whether it is becoming better or worse. Whatever the present situation is, it seems to me probable that dissatisfaction will increase as a result of rising educational and aspirational levels of workers as suggested by Bowles in the reviewed book (pp. 492, 498). Also, the strict hierarchical control to which workers are subjected contradicts the spirit of our time. Finally, previously satisfied workers may become dissatisfied from mere discussion of the subject, as happened in the Goldthrope experiment described by Andre Gorz in *The Capitalist System* (pp. 479–80). Workers' management or at least participation in decision-making has been suggested as a remedy; see, for instance, Hunnius (1973) and Jenkins (1973).

[8] See, however, pp. 273, 315, and a few others. In another radical book (Hunnius 1973), the reasons for the unions' attitudes are discussed and the unions are severely censured.

But these are a priori considerations, which may or may not be true. Strangely enough, Herbert Gintis, one of the contributors, has already decided that the "so-called socialist" worker is not any better off than his capitalist counterpart because of the similarity of their basic economic institutions (p. 277).

In any case, much of what the Soviet man gains as a worker he loses as a consumer. It would be possible to eliminate alienation in any society by the simple expedient of allowing each worker to produce whatever he wants, irrespective of consumer desires. Our authors do not suggest this extreme solution: in their ideal society (see Section IV below), workers' and consumers' preferences are reconciled by some unspecified mechanism, but their sympathy, like that of socialists in general, is for the worker. For the consumer they show little more than contempt (see below). But it is curious that in the university—one of our few organizations where at least some of the workers, namely, the professors, do make decisions—our radicals call for more power to students, that is, to consumers! It is also strange that our authors show little interest in, and no sympathy for, the one country—Yugoslavia—where workers have more decision power than in any other. It seems that Yugoslavia cannot be forgiven for her use of the market mechanism and of profits (pp. 483, 525–27).

Irrationality.—This big word means that our economy produces the wrong assortment of goods and services. To the extent that this assortment is determined by the existing distribution of income and wealth, I believe that our authors have a good point, as they have regarding the composition of government expenditures (although neither they nor I have proof that our views are shared by the majority of the American electorate). But here our ways part.

If in the old days (and even in days not so old) capitalism was blamed for unemployment and for the exploitation of workers (pp. 50, 71, 465–67), now it is accused of depriving workers of any joy from work and thus forcing them to become addicted to consumption as the only pleasure in life. This phenomenon is called "consumerism." It is treated with so much disgust (pp. 22, 369–70) that the old-fashioned exploitation might have been a lesser evil.

Consumerism exaggerates the importance of material things; the addicted workers buy millions of useless objects that they do not need and do not even enjoy (pp. 362, 376, 391, 404). Unfortunately, the lists of these useless objects and of the consumers who buy them are not provided. (Presumably our authors are innocent of these offenses.) It takes a remarkable intellectual arrogance, reminiscent of the old aristocratic contempt for the "lower orders," to hold this position (pp. 20–21, 268, 376). We may expect that after the revolution, the production of objects of which our authors disapprove will cease.

The mechanism that creates artificial desires to be satisfied by useless objects is advertising (pp. 283, 362, 369, 376, 378, 381).[9] But advertising is merely an instrument. The real cause of irrationality (and of many other capitalist ills) lies in the profit motive (pp. 90–91, 99–106, 274, 363, 383–86, 410–11). Goods are produced for profit, not for use.

A few years ago, a Soviet manager of a trucking firm was publicly castigated for using plan fulfillment as the criterion for ranking his shipments. The manager recognized his error and promised to mend his ways. But neither he nor his critics ever indicated how he, a mere trucking boss, could possibly rank his shipments in order of their marginal social utility, as his critics (implicitly) demanded.

The literature on the defects of profit maximization as an allocational device can fill a library, and yet through some 50 years of searching, the socialist countries have not discovered a better alternative.[10] Nor do our authors offer any. They are completely unconcerned with the problem of transforming a given social objective into microdecision rules.[11] It does not even occur to them that if their publishers, a good capitalist firm, had tried to maximize social welfare according to their own capitalist lights, rather than to make profits, they would not have published this book.

Inequality.—The facts are not in dispute: American income distribution is highly unequal, and that of wealth is even more so, as confirmed by several standard tables in the book. For some reason, none of the well-known tables by Simon Kuznets (1953) (partially reprinted in the *Historical Statistics of the United States*) showing a considerable improvement in income distribution between 1913 and 1948 is presented. But the book does contain a table (10-H, p. 445) indicating a more equal distribution of income in this country than in a number of others.

---

[9] In insisting on the effectiveness of advertising, our authors merely repeat what the advertising industry advertises about itself. Actually, as Schmalensee (1972) has discovered, there is not much reliable information about its effects; see, however, Taylor and Weiserbs (1972).

[10] Since income distribution in the socialist countries is closer to their ideal than ours is to our ideal, and since managers are confronted with parametric prices, profit maximization by socialist managers might be more appropriate than by capitalist managers, provided, of course, that prices were correctly set.

[11] In an amusing little book, Soviet engineer Antonov (1965) gives many examples of antisocial behavior by Soviet managers. His solution is not to induce them to behave altruistically, but to set such success indicators as would make them behave in a socially desirable manner while maximizing their own objective functions. The tone and the title of the book (*For All and for Oneself*) suggest the need for Adam Smith's "invisible hand."

The authors missed a good debating point by their failure to compare income distribution of capitalist and socialist countries: according to two recent studies, in the latter it is more equal.[12]

Granted that, the practical question is whether the effective functioning of the capitalist system requires as much inequality as is actually observed. Only social experimentation can definitely answer this question, but judging by the experience of other advanced capitalist countries, a considerable movement toward equality is not likely to ruin us. Rather unexpectedly, the book arrives at the opposite conclusion: "The capitalist mode of production is characterized by a serious conflict between income equality ... and economic efficiency .... A high degree of income equality could be attained in a capitalist society only at a very high cost in productive efficiency" (p. 128). A number of such passages are found in the book (pp. 127–28, 208, 249). Soon they will be quoted by business speakers. Imagine the delight of their audiences when they learn the source!

There are two common objections to the concentration of income and wealth: (1) if the rich have more, the poor have less, and (2) wealth conveys power. It is not quite clear why our authors should be concerned with the first objection. They emphasize the diminishing utility of income (pp. 284, 381) and the lack of connection between income and happiness (pp. 284, 362); they oppose economic growth (pp. 284, 371); they deplore the interest in material things and consumerism in general.[13] If higher incomes for the poor will merely increase their addiction to consumption without making them any happier, why bother about it at all?

It should be possible to eliminate the power of the rich by depriving them of their riches, but it would be much more difficult to diffuse their power. The absence of large stockholders would enhance the power of corporate managers, and the nationalization or public control of corporations would transfer the power to the government, even to one headed by a Nixon and worse, a familiar problem in the socialist countries.

Racism.—Again, the facts are not in dispute. Only the blind would deny the existence of racism in this country.

In my limited space, it is difficult to say much about this complex problem without repeating the trivial and the obvious. A narrow definition of racism would single out three countries: the United States, South Africa, and Rhodesia—all capitalist. But there is much less racism in capitalist Latin America, and particularly in Brazil. On the other hand, it is present in semisocialist India and in precapitalist Africa. If the definition of racism is to include the dominance by specific national or ethnic groups as well, there is no shortage of it in the Soviet Union, Rumania, Yugoslavia, and probably in China. Competition for jobs, particularly in periods of unemployment, should make capitalist countries more vulnerable. On the other hand, if capitalists are as determined to maximize profits as they are described in the book, why should they bother to discriminate? Indeed, The Netherlands, the most money-minded country in the seventeenth century, was also the most tolerant. The Dutch discovered that discrimination interfered with business.[14]

---

[12] The comparison of income distribution among different economic systems is, of course, a very complex and difficult task. Among other problems, nonmonetary incomes play a more important role in the socialist than in the capitalist countries. Nevertheless, Wiles and Markowski (1971) have found that income distribution is more equal in Poland than in the United Kingdom, and more equal in the latter than in the United States. Also, the distribution of Soviet nonagricultural income is more equal than the American one. Pryor (1972, 1973) reports that income distribution in socialist countries is more equal than in capitalist ones, even when agricultural income is included. In recent years, the position of Soviet peasants has improved considerably.

[13] Our authors' attitude to material wealth resembles that of the early Jesuits; see Harney (1941).

Perhaps it would be more fruitful to leave these generalities and inquire instead into the beneficiaries of discrimination.[15] Traditional wisdom, speaking through Gary Becker (1971), names white workers and black capitalists (such as there are). Edward Reich, in a refreshingly interesting paper on "The Economics of Racism" (pp. 313–21), by far the best in the book, blames white upper-income groups.[16] Paradoxically both may be right.

If white capitalists refuse to hire black workers, as Becker assumes, his conclusion follows. But most blacks in the labor force are employed. They are employed, however, in low-paid occupations.

Let us assume that all occupational differences between whites and blacks are caused by discrimination. When that is gone, the occupational structures will (eventually) become identical. Then the low-paid white workers will gain because of the upward movement of many of their black competitors, while the high-paid white workers will lose. If our assumption covers the distribution of wealth as well, black capitalists will gain at the expense of the white. Thus, it is the high-paid white workers and capitalists who gain from discrimination, as Reich suggests.

But all this after many years. In the meantime, the burden of equalization will fall on white semiskilled and even some unskilled workers, long before it reaches high-paid professionals and capitalists. Thus, the current, if only temporary, beneficiaries of discrimination are neither at the top nor quite at the bottom of income distribution, but somewhere below the median, as confirmed, more or less, by the intensity of racial prejudice among these groups.[17]

---

[14] Jews were allowed by Cromwell to return to England, and were emancipated in France by the Great Revolution, as they were in other countries with the development of capitalism.

[15] In this discussion, the gain or loss from discrimination is merely relative. The country as a whole would, of course, gain from equal treatment of all citizens because of the resulting improvement in the use of resources.

[16] Reich found a correlation (in a multiple regression) between the degree of income concentration among whites and racism defined as the ratio between black and white median incomes by metropolitan areas. Since only a summary of his work (a doctoral dissertation) is presented in the book, I cannot comment on his data and methods.

But if Reich is to be commended for a useful contribution, he may be reproached for failing to explain how the alleged interests of the capitalists as a class are promoted by everyday actions of individual capitalists. Are the latter so devoted to their class that they are ready to forego larger profits to be derived from hiring blacks? Reich might have also said more about racist behavior of white workers (and of some unions), unless he thought that their motivation was too obvious to require a lengthy explanation.

Sexism.—Since the problem is similar to racism, there is no need to go over the arguments again. But it would be only fair to record that in socialist countries women's participation in the labor force is greater than it is in ours. They are active in medicine, engineering, science, construction, administration, etc. But the performance of two sets of duties is anything but easy.[18]

Imperialism.—If it is defined as "the internationalization of capitalism" (pp. 408, 417), then everything is conveniently settled by definition. But if imperialism means a directed expansion of an economic system beyond the boundaries of a particular country that intervenes and acquires control over other areas, as the book further elaborates (p. 408), then between the American and Soviet imperialisms there is little to choose. To preserve its position and power at home, one ruling group invades Vietnam; the other, Hungary and Czechoslovakia—not to mention their other aggressions. Whether capitalism is more likely to pursue imperialist policy than socialism is a subject on which many arguments and counterarguments can be made without settling much, particularly in a limited space.[19] Let me turn instead to Weisskopf's two papers, "United States Foreign Investment: An Empirical Survey" (pp. 426–35) and "Capitalism and Underdevelopment in the Modern

[17] That these conclusions are strongly affected by the nature of assumptions made can be seen by comparing them with Barbara Bergmann's (1971) findings. She assumed that the elimination of discrimination would reshuffle workers only in each educational group and found that the greatest losers would be low-educated whites.

[18] Several years ago, the Soviet literary journal *Novyĭ Mir* [New world] published a story named "Nedelia" [Week] in the form of a diary by a young married professional woman, with two children, describing her activities during 7 days. According to her, holding a full-time job and taking care of her family was extremely difficult. If more American women enter the labor force and earn incomes, would not consumerism increase? See the story about the female employees of the telephone company on pp. 20–21 of the book under review.

[19] Those who are sure that socialist countries are not imperialist might ponder on the likely Soviet behavior if the Soviet supply of oil were threatened.

World" (pp. 442–57). Both look like interesting articles with analytical and empirical content, but both turn out to be little better than lawyer's briefs.

Thus, Table 10-B (p. 429) is supposed to demonstrate that in 1950–69 the rate of return on American foreign investment—13.3 percent—was much higher than the overall one of 7.7 percent. But these rates are net of taxes. A simple recalculation based on Weisskopf's own assumptions reveals that the pretax rates differ little; they are 14.8 and 14.2 percent, respectively.[20] Profit rates are notoriously inexact and subject to manipulation, so I would rather abstain from conclusions, but what was the purpose of presenting this table? To show that income from foreign investment is taxed lightly?

Table 10-G (p. 444) is supposed to illustrate the uneven character of capitalist development. Indeed, among the nonsocialist countries, the ratio of per capita income of the rich to the poor is 12 to one. But for socialist countries (here they are called "socialist"), this ratio is as high as 10 to one, certainly high enough to proclaim a new "Law of Uneven Socialist Development." It seems that India and Egypt are placed in the nonsocialist category. As both regard themselves as semisocialist, perhaps the honor of their company should be shared by both categories. I wonder what the two ratios will be then.

Still another table, this time 10-I (p. 455), shows a widening gap between the per capita incomes of rich and poor nonsocialist countries. But the fact that population growth in the poor countries is faster than in the rich (as given in that table) is not mentioned in the text, although, in the spirit of Weisskopf's discussion, population growth hardly contributes to the growth of output.[21] It seems that population growth is still a forbidden subject among many radicals. Now that even China has embarked on population control, should not this taboo be lifted?

According to Weisskopf, the poor capitalist countries are doomed whatever they do. They gain little from investment because the latter merely increases labor productivity instead of alleviating unemployment (pp. 449–50), as if it would be difficult to make investment less labor-saving if only someone was willing to pay the cost. Industrialization merely raises incomes of industrial workers, already above the average, and thus increases inequality (p. 456); evidently low-earning peasants never move into industrial jobs. His own data (Table 10-H, p. 445) show that advanced capitalist countries enjoy greater income equality than the underdeveloped ones. I wonder how the former have ever managed to achieve this?

---

[20] Without examination, I took Weisskopf's estimate of the average tax rate on foreign investment as 10 percent. For the overall rate I used 45.6 percent. The gross profit rates so obtained are simple averages calculated from Weisskopf's data. Since profit from foreign investment constitutes a small part of total profit, the small difference in the pretax profit rates stated in the text implies considerable difference between foreign and domestic rates. It is puzzling that Weisskopf used the overall, rather than the domestic, rate for his comparison.

[21] Actually, the total income of poor countries was growing faster (at 4.6 percent per year) than that of the rich (4.4), but the population of the former was growing at 2.4 percent, as compared with only 1.3 percent for the latter.

IV

Poor Capitalism.—As the old saying goes, a glass of water can be described either as half full or as half empty. Our authors never tire of playing this game. Poor capitalism! It is damned if it does, and it is damned if it does not. Thus, a change in the relative positions of white- and of blue-collar workers is described as a loss for the former rather than as a gain for the latter (pp. 180, 256–57). Federal policies for economic stability and growth are blamed for leading "to the survival of inefficient business, and hence, in the long run to the need for more subsidies" (p. 197). What a surprising piece of social Darwinism! The United States is scolded for imposing a brain drain on the poorer countries (p. 447). But if we forbade the entrance of their nationals, the same authors would accuse us of racism, since many of these immigrants are nonwhite. A quotation from *The Communist Manifesto* deplores the fact that "labor of men [is] superseded by that of women" (p. 71)—a rather strange idea for sympathizers with Women's Liberation. Even minimum-wage legislation is bad because "it can serve the interests of organized labor at the cost of overpricing and hence underutilizing unskilled labor" (p. 453). Milton Friedman would agree.

Our educational system is severely criticized for, among other things, preparing people for productive jobs, for being "more or less firmly tailored to the needs of 'economic rationality'" (p. 124). Shall we train economically useless graduates instead? Compulsory education is described as being "basically coercive.... In many parts of the country, schools were literally imposed upon the workers" (p. 221). Why not repeal it, then?

These are a few examples. I have run out of space.

The promised land.—To condemn capitalism is easy, to present a superior practical alternative is difficult, particularly for our authors: their rejection both of the market and of bureaucracy as instruments for resource allocation would require a truly ingenious substitute. Unfortunately, none is offered.

The "Visions of a Socialist Alternative" is given only 20 pages at the end of the book, plus a number of scattered suggestions. Here is a brief summary.

People will live in communities based on "geographical contiguity" (p. 527). The communities consisting of a "variety of functional groups"

(p. 527) will control productive wealth and make economic decisions. They will be "unoppressive, nonexploitative ... where individuals are encouraged to lead creative lives" (p. 347). Their members will be motivated by "a cooperative ethic of recognizing people's responsibility to each other" (p. 520). They will work for the joy of working and not for income or profit (pp. 520–28). This wonderful transformation in human behavior will come about because people are not "inherently greedy, acquisitive, selfish, competitive or aggressive" (p. 4). "Changes in the environment can interact with changes in the individual to usher a new era of human cooperation" (p. 5).

The economic mechanism that will be used to make decisions within the communities, and—more important—to allocate resources and to organize exchange among them, is not described. But with human nature so good and pliable (in the right hands, of course), who needs formal organizations, markets, prices, plans, and all other economic paraphernalia?

So the end result is just another utopia, recognized by the authors as such (pp. 392, 530). It is an old-fashioned anarchist utopia that would delight Kropotkin and Proudhon (and Furier), but hardly please Marx, if he remained true to his own spirit. In its treatment of economic problems, it is not superior to Thomas More's original creation, and it is greatly inferior to Edward Bellamy's *Looking Backward* ([1888] 1960), now nearly 100 years old. And Bellamy was not even an economist!

There is no harm in describing utopias if one does not take them seriously. But what is the use of criticizing capitalism, or any other existing economic system, in a supposedly scholarly and analytical manner, by comparing it with an ideal, which can be made as wonderful as the authors' imagination allows? Surely more effective methods can be found. The ineptitude shown by the contributors and the editors (well-trained young economists of known ability) merely damages their own cause: it makes capitalism look better than it is. Instead of winning converts, they are more likely to repel even those who have no love for capitalism and are searching for better alternatives.

References

Antonov, O. K. *Dlia Vsekh i dlia Sebia* [For all and for oneself]. Moscow: Ekonomika, 1965.

Becker, G. S. *The Economics of Discrimination.* 2d ed. Chicago: Univ. Chicago Press, 1971.

Bellamy, E. *Looking Backward.* New York: New American Library, 1960.

Bergmann, B. "The Effect on White Incomes of Discrimination in Employment." *J.P.E.* 79 (March/April 1971): 294–313.

Edwards, Richard C.; Reich, Michael; and Weisskopf, Thomas B. *The Capitalist System: A Radical Analysis of American Society.* Englewood Cliffs, N.J.: Prentice-Hall, 1972.

Harney, M. P. *The Jesuits in History*. New York: American Press, 1941.

Hunnius, C.; Garson, G. D.; and Case, J., eds. *Workers' Control: A Reader on Labor and Social Change*. New York: Vintage, 1973.

Jencks, C., et al. *Inequality: A Reassessment of the Effect of Family and Schooling in America*. New York: Basic, 1972.

Jenkins, D. *Job Power: Blue and White Collar Democracy*. Garden City, N.Y.: Doubleday, 1973.

Kuznets, S. *Shares of Upper-Income Groups in Income and Savings*. New York: Nat. Bur. Econ. Res., 1953.

Osipov, G. V., ed. *Industry and Labor in the U.S.S.R.* London: Tavistock, 1966.

Pryor, F. L. "The Distribution of Nonagricultural Labor Incomes in Communist and Capitalist Nations." *Slavic Rev.* 31 (September 1972): 639–50.

———. *Property and Industrial Organization in Communist and Capitalist Nations*. Bloomington: Indiana Univ. Press, 1973.

Schmalensee, R. *The Economics of Advertising*. Amsterdam: North-Holland, 1972.

Taylor, L. S., and Weiserbs, D. "Advertising and the Aggregate Consumption Function." *A.E.R.* 62 (September 1972): 642–55.

U.S., Bureau of the Census. *Historical Statistics of the United States: Colonial Times to 1957*. Washington, D.C.: Superintendent of Documents, 1960.

Wiles, P. J. D., and Markowski, S. "Income Distribution under Communism and Capitalism: Some Facts about Poland, the UK, the USA, and the USSR." *Soviet Studies* 22 (January 1971): 344–69; (April 1971): 487–511.

Work in America: Report of a Special Task Force to the Secretary of Health, Education, and Welfare. Washington: Government Printing Office, 1972.

Zdravomyslov, A. G.; Roz͡hin, V. P.; and I͡adov, V. A., eds. *Man and His Work*. White Plains, N.Y.: International Arts and Science Press, 1970.

Book Reviews

Opening Up the Suburbs. By Anthony Downs.
New Haven, Conn.: Yale University Press, 1973. Pp. ix + 219. $7.95.

This is an essay in persuasion rather than scholarship. It is addressed to middle-class suburban Americans and attempts to persuade them to reconsider their opposition to opening suburban communities to poor, and especially black, residents of central cities. It lacks the documentation and careful analysis of a scholarly book, but there is a great deal of careful scholarship lurking in the background.

Whatever one thinks of its conclusions, it is a remarkable book. It is calm and conciliatory on a subject on which almost no one is calm and conciliatory. Downs understands, sympathizes with, and, to a considerable extent, shares the goals and anxieties of white suburbanites. But he calmly and firmly insists that the goals and rights of the central city poor be given increased weight. He avoids the temptation to score cheap points against extremists on both sides. Yet there is no room for doubt about where he stands.

Rapid postwar metropolitan expansion has inevitably spilled across boundaries of central city governments. For a variety of reasons, the first urban migrants into many suburban jurisdictions were predominantly people of above average incomes. By the use of zoning, other land-use controls, hostility, and occasionally violence, low-income and, especially, black central city residents have been denied entry into many relatively affluent suburbs. Many of the early chapters in Downs's book lay out the consequences of this situation for central city residents. It has impaired their access to jobs in suburbs, where most metropolitan employment growth has occurred. It has meant that their housing has been of poorer quality and higher price than otherwise. It has limited their access to high-quality local public services, especially education. It has prevented them from escaping antisocial behavior, especially crime, which is concentrated in central cities. And, Downs believes, it has deprived them of access to middle-class models of achievement, orderly living, and good citizenship.

Downs wants to open up the suburbs. He does not necessarily favor racial and income integration at the level of a small neighborhood because he believes that most people want to live in the immediate vicinity of others like them in many ways. But he wants integration at the level of commuting areas for large suburban job centers, school districts, and jurisdictions of other local government services. Furthermore, he is anxious to permit suburbs to preserve middle-class predominance. He favors (p. 140) an upper limit of 25 percent of low and moderate-income schoolchildren in any school where the percentage is now below that level. Finally, he recognizes the advantage of simultaneity. If a single suburb lowers its barriers, it might be overwhelmed with central city migrants. That is less likely if all the suburbs in a metropolitan area do so at about the same time, since the nonpoor outnumber the poor by about 10 to 1 in most metropolitan areas.

How does Downs intend to open up the suburbs? Given the hostility to the idea in the suburbs, in Washington, and, increasingly, among black leaders, it is easy to dismiss the idea as that of an ivory tower idealist. But that would underestimate Downs. He frankly admits he is not sanguine. But he makes a variety of interesting proposals which cannot be summarized here. Some would provide central city residents with incentives to suburbanize and some would provide suburbanites with incentives to accept them. Some could be carried out by local, some by state, and some by national governments.

I want to comment further on the simultaneity issue which Downs does not emphasize. The danger of being overwhelmed by black and poor people is surely sufficient to prevent individual (or small groups of) surburban communities from unilaterally lowering barriers. Thus, action must come at the metropolitan area level or at a higher level of government. This is important not only to suburbanites but also to central city residents. Downs emphasizes that if a single suburb opens up it may become a suburban slum with all the problems of central city slums, thus frustrating the efforts of central city residents to improve their lives. More than that, if Downs's 25 percent quota were put into effect without simultaneously opening up all suburbs consistently with the quota it could be disastrous. For example, Prince Georges County, northeast of Washington, is now inhabited by large numbers of black former District residents, whereas Montgomery County and the Virginia suburbs are still largely closed to blacks. If the 25 percent quota had been in effect in Prince Georges suburbs, without effectively opening the other suburbs, it would have largely closed off the one escape from the District ghetto that has been open to blacks. Since we have practically no metropolitan area-wide governments and since metropolitan areas cross state lines, I believe simultaneity can only be achieved by firm federal action.

Among the most controversial aspects of the book is the emphasis Downs places on preserving middle-class dominance of suburban communities. Some will regard this as disguised racism. But a fairer interpretation would be to regard it as a tactical consideration. Many suburbanites feel threatened by the chaos of central city ghettos, and their fears are by no means entirely irrational. Realistically, there is no way to integrate the suburbs without lowering suburbanites' anxiety levels. Downs is anxious to do that while still achieving the basic goal of suburban integration. More controversially, he believes that arrangement would be best for the poor. He believes, along with some education specialists, that it would improve the school performance of blacks and poor without greatly impairing that of white middle-class suburban children. And he believes that crime rates and other forms of antisocial behavior would be lessened. Finally, he believes that increases in income and reductions in housing costs that suburban integration would enable the poor to achieve would lessen antisocial behavior. Better understanding of these issues is crucial. If the optimistic contention is correct, white suburbanites stand to lose little from integration. If, on the other hand, integration would make their schools nearly as chaotic as many in the ghettos and their streets as dangerous, then there is almost no hope that they can be persuaded to lower the barriers.

No one will agree with everything in this book, and I would oppose some of his policy proposals. But I find its basic proposition persuasive. In a society in which economic and political opportunities were widely distributed, small exclusionary suburbs would be hardly more objectionable than exclusionary country clubs, and they would have significant advantages. But in a society in which almost every aspect of public and private life has been organized to oppress the black minority for so long, continued use of the police power to exclude them from residence in large parts of metropolitan areas is intolerable.

EDWIN S. MILLS

Princeton University

The Demand for State and Local Government Employees. By RONALD G. EHRENBERG. Lexington, Mass.: D. C. Heath & Co., 1972. $15.00.

The growing importance of economic allocative decisions made within the public sector, particularly at the state and local level, has called forth a growing literature that tries to analyze either the time-series or state cross-section variation in expenditure patterns by various functional categories (type of service). Ehrenberg's book is one of several recent studies that make important improvements in the theoretical and econometric foundation of the analysis of governmental decision making. The focus of his book is the derived demand for labor in 11 specific governmental service categories. His analysis is based on a combination of the available time-series and state cross-section data from which he estimates wage and income elasticities of demand together with the effects of federal grants to states and certain demographic variables. In a final chapter he uses his estimates as the basis for a set of simulation experiments in which he considers the employment effects of various patterns of wage changes and of a spending program designed for job creation (a public employment program).

In Ehrenberg's model the typical state consumer has preferences defined over privately and publicly provided goods and services. The assumption of a separable utility function permits Ehrenberg to treat the allocation decision in two stages. First, the allocation between public and private sector expenditure is made on the basis of the total resource constraint and the relative price of the two sectors' services. In a second step the public sector expenditure is allocated among functional categories on the basis of their relative prices. Public sector goods and services are provided by a constant-returns production technology that is common to all states.

The basic structure of the Ehrenberg model is similar to that of an independent study completed about the same time by Borcherding and Deacon (1972), and the two studies invite comparison. Borcherding and Deacon postulate demand functions for publicly provided goods and services, which are in turn produced using a constant-returns production technology employing capital and labor. They focus on the demand for the services rather than the derived demand for labor, but the conversion is a simple exercise. This structure is familiar as the Allen (1938) model of a demand-limited competitive industry. From their common skeleton the models are specialized in different ways. Borcherding and Deacon provide some refinement of the social choice aspects of the model. They assume constant-elasticity demand functions for the governmental services. Ehrenberg, in his theoretical development, uses the Klein-Rubin-Stone linear expenditure framework, although he largely abandons it in the econometric work in favor of constant-elasticity demand functions. On the supply side, Borcherding and Deacon assume Cobb-Douglas production functions, solving for the self-dual cost function, then dodging the nasty problems of constructing capital service and rental price series by assuming that there is no variation in the capital service price across states. Ehrenberg also tries to dodge the capital problems (there may be no neat way to meet the problems with existing data) by assuming that capital and labor are used in fixed proportions. This seems an unfortunate choice for two reasons. First, the wealth of production studies suggests that unity rather than zero is the better guess for the elasticity of substitution between capital and labor. Second, Ehrenberg's choice leads him into a minor specification error. The supply price for public services is the weighted sum of the capital and labor prices. Ehrenberg uses the wage rate as a proxy, but in logs the two series are not collinear even if the capital price is invariant across states. This problem does not arise in the Cobb-Douglas case.

The wage rate provides the only price variability in the model, but simultaneous labor-quality variation could confuse the price information. Ehrenberg recognizes the problem, and in fact presents a model to show the possibility that wage changes in the public sector could lead to changes in both labor-quality and vacancy rates. As a practical matter, however, he uses in his regressions the uncorrected state average wage for each category; hence, the quality-variation problem remains. He also carries along a parallel set of regressions in which he uses a relative wage series, the category wage divided by the public sector wage index. These results, though similar to those resulting from the uncorrected wage series, are generally weaker.

Ehrenberg's data consist of 12-year time series of state cross-section observations on employment and the average wage rate for each of 11 functional categories of governmental services. In addition he uses state personal income, federal grants to states, and a set of demographic variables in his estimating equations. The results that he focuses on and uses for the simulation experiments are based on the pooled time-series and cross-section data, although he also presents results for individual-year cross sections and for individual-state time series. One important conclusion that emerges is that the yearly cross section results tend to be unstable. Many previous studies, including Borcherding and Deacon's, are based on such single-year cross sections and hence should be used cautiously.

By and large Ehrenberg's results indicate significantly negative price response in the public sector. If the labor-quality variation is ignored, his labor-demand elasticities can be interpreted (approximately) as labor's share in total cost multiplied by the elasticity of demand for the particular service. For the year 1962 for the functional categories that are common to the two studies, the Ehrenberg-derived demand elasticities are generally smaller in magnitude than those found by Borcherding and Deacon. Part of the difference should be expected as the result of the differences in their choice of substitution parameter. Another part of the difference, however, may result from the use of the uncorrected wage-rate series by Ehrenberg in contrast with the quality-corrected wage series used by Borcherding and Deacon.

The econometric work is generally careful and explores a variety of approaches to the estimation of the model. Because of the size of the model, Ehrenberg did not fully exploit the linear-expenditure model's features. In particular there is no examination of cross elasticities of demand within the public sector. I found it puzzling that he ignored the recursive nature of his structural model and used two- and three-stage least squares when ordinary least squares and seemingly unrelated techniques would have sufficed.

Ehrenberg's study will be of interest to students of public sector behavior. The simulation experiments have a number of conclusions relevant for manpower planning and union activity in the public sector. Finally, it represents a reasonable starting point for someone willing to work on problems of the dynamics of choice and the dynamics of public sector capital decisions whose presence is signaled by the instability of the yearly cross-section estimates.

RICHARD W. PARKS

University of Washington

References

Allen, R. G. D. *Mathematical Analysis for Economists.* London: Macmillan 1938.

Borcherding, T. E., and Deacon, R. T. "The Demand for Services of Non-Federal Governments." *A.E.R.* 62 (December 1972): 891–901.

STATEMENT OF OWNERSHIP, MANAGEMENT AND CIRCULATION (as required by Act of August 12, 1970. Section 3685 Title 39, United States Code). 1 Title of publication. Journal of Political Economy, 2. Date of filing. August 20, 1974; 3 Frequency bi-monthly; 4. Location of known office of publication: 5801 Ellis Avenue, Chicago, Illinois 60637; 5. Location of the headquarters or general business offices of the publishers: same address; 6. Names and addresses of publisher, editor, and managing editor: Publisher, University of Chicago Press, 5801 Ellis Avenue, Chicago, Illinois 60637, Editor, George J Stigler, Harry G Johnson, University of Chicago, 5801 Ellis Avenue, Chicago, Illinois 60637; 7. Owner University of Chicago Press, 5801 Ellis Avenue, Chicago, Illinois 60637; 8. Known bondholders, mortgagees, and other security holders owning or holding 1 percent or more of total amount of bonds, mortgages, or other securities: none; 9 For optional completion by publishers mailing at regular rates: not applicable; 10. For completion by nonprofit organizations authorized to mail at special rates, the purpose, function, and nonprofit status for Federal income tax purposes have not changed during preceding 12 months; 11. (a) Total number of copies printed. Average number of copies each issue during preceding 12 months, 7,633; Single issue nearest filing date, 8,000, (b) Paid circulation Sales through dealers and carriers, street vendors and counter sales, none; Mail subscriptions, average number of copies each issue during preceding twelve months, 6,834. Single issue nearest filing date, 7,038; (c) Total paid circulation. Average number of copies each issue during preceding twelve months, 6,834; Single issue nearest filing date, 7,038; (d) Free distribution. Samples, complimentary, and other free copies, Average number of copies each issue during preceding twelve months, 137, Single issue nearest filing date, 137. Copies distributed to news agents, but not sold, none; (e) Total distribution (sum of c and d): Average number of copies each issue during preceding twelve months, 6,971, Single issue nearest filing date, 7,175; (f) Office use, leftover, unaccounted, spoiled after printing: Average number of copies each issue during preceding twelve months, 662; Single issue nearest filing date, 825; (g) Total. Average number of copies each issue during preceding twelve months, 7,633; Single issue nearest filing date, 8,000.

I certify that the statements made by me above are correct and complete. (Signed) Jean W Sacks, Assistant Director

Index to Volume 82

Articles and Notes

Amacher, Ryan C., and Hodgson, John S. Purchasing-Power Parity Theory and Economic Reform in Yugoslavia 809

Auernheimer, Leonardo. The Honest Government's Guide to the Revenue from the Creation of Money 598

Bailey, Martin J. Progressivity and Investment Yields under U.S. Income Taxation 1157

Barro, Robert J. Are Government Bonds Net Wealth? 1095

Batra, Raveendra N., and Ullah, Aman. Competitive Firm and the Theory of Input Demand under Price Uncertainty 537

Becker, Gary. A Theory of Marriage: Part II S11

Becker, Gary. A Theory of Social Interactions. 1063

Benham, Lee. Benefits of Women's Education within Marriage S57

Berglas, Eitan, and Razin, Assaf. "Preferences, Separability, and the Patinkin Model": A Comment 199

Blinder, Alan S. The Economics of Brushing Teeth 887

Bloch, Harry. Advertising and Profitability: A Reappraisal 267

Bomball, Mark R. See *Primeaux, Walter J., Jr.*

Boskin, Michael J. A Conditional Logit Model of Occupational Choice. . 389

Branch, Ben. Research and Development Activity and Profitability: A Distributed Lag Analysis 999

Brovender, Shlomo. On the Economics of a University: Toward the Determination of Marginal Cost of Teaching Services 657

Brown, Gardner, Jr. An Optimal Program for Managing Common Property Resources with Congestion Externalities. 163

Camacho, Antonio, and Sonstelie, Jon. Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparison of Utilities: A Note 607

Cell, Donald C. Policy Influence without Policy Choice 1017

Chiswick, Barry, and Mincer, Jacob. Time-Series Changes in Personal Income Inequality in the United States from 1939 with Projections to 1985: Reply 1033

Comay, Yochanan; Melnik, Arie; and Subotnik, Abraham. Bargaining, Yield Curves, and Wage Settlements: An Empirical Analysis 303

Cyert, Richard M., and DeGroot, Morris H. Rational Expectations and Bayesian Analysis 521

Darby, Michael R. Consumers' Durable Goods, Human Capital, and the Pareto Optimality of the Lange-Lerner State 180

Dasgupta, Partha, and Stiglitz, Joseph E. Benefit-Cost Analysis and Trade Policies 1
David, Paul A., and Scadding, John L. Private Savings: Ultrarationality, Aggregation, and "Denison's Law" 225
De Alessi, Louis. Managerial Tenure under Private and Government Ownership in the Electric Power Industry 645
Deardorff, Alan V. Factor Proportions and Comparative Advantage in the Long Run: Comment 829
DeGroot, Morris H. See *Cyert, Richard M.*
Denslow, David, Jr., and Schulze, David. Optimal Replacement of Capital Goods in Early New England and British Textile Firms: A Comment . 631
Dick, G. William. Authoritarian versus Nonauthoritarian Approaches to Economic Development 817
Domar, Evsey D. Poor Old Capitalism: A Review Article 1301
Duncan, Otis Dudley. Comment S109
Eatherly, Billy J. Drug-Law Enforcement: Should We Arrest Pushers or Users? 210
Edel, Matthew, and Sclar, Elliott. Taxes, Spending, and Property Values: Supply Adjustment in a Tiebout-Oates Model 941
Eysenbach, M. L. Voucher Plans, Voting Models, and the Efficiency of Local School Finance 863
Feldstein, Martin. Social Security, Induced Retirement, and Aggregate Capital Accumulation 905
Freiden, Alan. The United States Marriage Market S34
Friedman, Milton. A Bias in Current Measures of Economic Growth . . . 431
Gandolfi, Arthur E. Stability of the Demand for Money during the Great Contraction—1929–1933 969
Gisser, Micha, and Jonas, Paul. Soviet Growth in Absence of Centralized Planning: A Hypothetical Alternative 333
Goldsmith, Oliver Scott. Market Allocation of Exhaustive Resources . . . 1035
Goode, William J. Comment: The Economics of Nonmonetary Variables . S27
Gordon, Scott. Frank Knight and the Tradition of Liberalism 571
Gould, J. R., External Economies and the Production Frontier 1239
Grant, Dwight. The Staple Theory and Its Empirical Measurement . . . 1249
Griffin, James M. An Econometric Evaluation of Sulfur Taxes 669
Griliches, Zvi. Comment S219
Gronau, Reuben. Wage Comparisons—a Selectivity Bias 1119
Grubel, Herbert G. Taxation and the Rates of Return from Some U.S. Asset Holdings Abroad, 1960–1969 469
Gwartney, James, and Haworth, Charles. Employer Costs and Discrimination: The Cost of Baseball 873
Hamermesh, Daniel S., and Soss, Neal M. An Economic Theory of Suicide . 83
Harrison, Bennett. Ghetto Employment and the Model Cities Program . . 353
Hashimoto, Masanori. Economics of Postwar Fertility in Japan: Differentials and Trends S170
Hause, John C. The Risk Element in Occupational and Educational Choices: Comment 803
Haworth, Charles. See *Gwartney, James.*
Heckman, James J. Effects of Child-Care Programs on Women's Work Effort S136
Henderson, J. V. Optimum City Size: The External Diseconomy Question 373
Hodgson, John S. See *Amacher, Ryan C.*
Holmes, James M. A Test of the Permanent-Income Hypothesis: Comment 185
Howitt, P. W. Stability and the Quantity Theory 133

Ijiri, Yuji, and Simon, Herbert A. Interpretations of Departures from the Pareto Curve Firm-Size Distributions 315
Jacobs, Rodney L. Estimating the Long-Run Demand for Money from Time-Series Data 1221
Jarvis, Lovell S. Cattle as Capital Goods and Ranchers as Portfolio Managers: An Application to the Argentine Cattle Sector 489
Johnson, Elizabeth. John Maynard Keynes: Scientist or Politician? 99
Johnson, George E., and Stafford, Frank P. Lifetime Earnings in a Professional Labor Market: Academic Economists 549
Jonas, Paul. See *Gisser, Micha*
Jones-Lee, Michael. The Value of Changes in the Probability of Death or Injury 835
Khan, Mohsin S. The Stability of the Demand-for-Money Function in the United States 1901–1965 1205
Koeune, Jean-Claude. On Some Unhappy Consequences of Professor Liebhafsky's Consumers' Demand for Psychiatric Help 1049
Koo, Anthony Y. C., and Schmidt, Peter. Cognitive Range in the Theory of Revealed Preference 174
Kouri, Pentti J. K., and Porter, Michael G. International Capital Flows and Portfolio Equilibrium 443
Laumas, Prem S. A Test of the Permanent-Income Hypothesis: Reply . . 192
Layard, Richard, and Psacharopoulos, George. The Screening Hypothesis and the Returns to Education 985
Lee, Joong-Koon. The Real Stock of Money and Velocity in a Wartime Economy: A Comment 1281
Leibowitz, Arleen. Home Investments in Children S111
Leigh, Arthur H. Frank H. Knight as Economic Theorist 578
Lewis, H. Gregg. Comments on Selectivity Biases in Wage Comparisons . 1145
McGuire, Martin. Group Segregation and Optimal Jurisdictions 112
Machlup, Fritz. Proxies and Dummies 892
McLure, Charles E., Jr. A Diagrammatic Exposition of the Harberger Model with One Immobile Factor 56
Makinen, G. E. The Real Stock of Money and Velocity in a Wartime Economy: A Reply 1285
Maurizi, Alex. Occupational Licensing and the Public Interest 399
Mayer, Wolfgang. Short-Run and Long-Run Equilibrium for a Small Open Economy 955
Meade, James E. A Note on Border-Tax Adjustments 1013
Melnik, Arie. See *Comay, Yochanan.*
Miller, H. Laurence, Jr. On Killing off the Market for Used Textbooks and the Relationship between Markets for New and Secondhand Goods . . 612
Mincer, Jacob. See *Chiswick, Barry.*
Mincer, Jacob, and Polachek, Solomon. Family Investments in Human Capital: Earnings of Women S76
Mishan, E. J. What Is the Optimal Level of Pollution? 1287
Montmarquette, Claude. A Note on Income (Labor) Inequality: Income Tax Systems and Human Capital Theory 620
Mueller, Dennis C.; Tollison, Robert D.; and Willett, Thomas D. On Equalizing the Distribution of Political Income 414
Mussa, Michael. Tariffs and the Distribution of Income: The Importance of Factor Specificity, Substitutability, and Intensity in the Short and Long Run 1191
Nelson, Phillip. Advertising as Information 729

Nerlove, Marc. Household and Economy: Toward a New Theory of Population and Economic Growth . . . S200
Nichols, Donald A. Some Principles of Inflationary Finance . . . 423
Oakland, William H. Public Goods, Perfect Competition, and Underproduction . . . 927
Orosel, Gerhard O. A Paradoxon of the Market Mechanism . . . 202
Park, Rolla Edward. How Analysis Affects Decision-Makers' Views: A Bayesian Model . . . 1041
Parsons, Donald O. The Cost of School Time, Foregone Earnings, and Human Capital Formation . . . 251
Parsons, Donald O. See *Schwartz, Aba.*
Patinkin, Don. A Note on the Stability of Metzler's Phase Diagram . . . 641
Peles, Yoram. A Note on Equilibrium in Monopolistic Competition . . . 626
Piron, Robert. Utility Maximization Sufficient for Competitive Survival: Comment . . . 654
Pissarides, Christopher A. Risk, Job Search, and Income Distribution . . . 1255
Polachek, Solomon. See *Mincer, Jacob.*
Porter, Michael G. See *Kouri, Pentti J. K.*
Primeaux, Walter J., Jr., and Bomball, Mark R. A Reexamination of the Kinky Oligopoly Demand Curve . . . 851
Psacharopoulos, George. See *Layard, Richard.*
Razin, Assaf. See *Berglas, Eitan.*
Roll, Richard. Rational Response to the Money Supply . . . 587
Rosen, Sherwin. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition . . . 34
Rosen, Sherwin. Comment . . . S164
Rosenberg, Nathan. Adam Smith on Profits—Paradox Lost and Regained . . . 1177
Rosenberg, Nathan. Karl Marx on the Economic Role of Science . . . 713
Rothschild, Michael. Searching for the Lowest Price When the Distribution of Prices Is Unknown . . . 689
Sandmo, Agnar. Investment Incentives and the Corporate Income Tax . . . 287
Santoni, Gary. See *Van Cott, T. Norman.*
Saxonhouse, Gary R. Comment . . . S195
Scadding, John L. See *David, Paul A.*
Schmidt, Peter. See *Koo, Anthony Y. C.*
Schultz, Theodore W. The High Value of Human Time: Population Equilibrium . . . S2
Schulze, David. See *Denslow, David, Jr.*
Schwartz, Aba, and Parsons, Donald O. Time-Series Changes in Personal Income Inequality in the United States from 1939 with Projections to 1985: A Comment . . . 1027
Sclar, Elliott. See *Edel, Matthew*
Sharir, Shmuel, and Weiss, Yoram. The Role of Absolute and Relative Wage Differentials in the Choice of a Work Activity . . . 1269
Simon, Herbert A. See *Ijiri, Yuji.*
Sonstelie, Jon. See *Comacho, Antonio.*
Soss, Neal M. See *Hamermesh, Daniel S.*
Stafford, Frank P. Comment . . . S132
Stafford, Frank P. See *Johnson, George E.*
Stiglitz, Joseph E. The Cambridge-Cambridge Controversy in the Theory of Capital; A View from New Haven: A Review Article . . . 8
Stiglitz, Joseph E. See *Dasgupta, Partha.*
Subotnik, Abraham. See *Comay, Yochanan.*

Thompson, Earl A. Taxation and National Defense 755
Tollison, Robert D. See *Mueller, Dennis C.*
Ullah, Aman. See *Batra, Raveendra N.*
Van Cott, T. Norman, and Santoni, Gary. Friedman versus Tobin: A Comment 883
Vining, Daniel R., Jr. Stability in the Pareto Law and the Popular Folklore: A Comment 1277
von Furstenberg, George W. The Interrelation between Labor and Capital Components of Racial Income Differences. 152
Wallace, T. Dudley. Comment S54
Warren-Boulton, Frederick W. Vertical Control with Variable Proportions . 783
Weinstein, Robert. Human Migration: A Survey of Preclassical Literature . 433
Weiss, Yoram. See *Sharir, Shmuel.*
Welch, Finis. Comment S72
Willett, Thomas D. See *Mueller, Dennis C.*
Williamson, Jeffrey G. Optimal Replacement of Capital Goods in Early New England and British Textile Firms: Reply 638

Reviews

Ashenfelter, Orley, and Rees, Albert, eds. *Discrimination in Labor Markets.* By Ronald G. Ehrenberg 1053
Boulding, Kenneth E. *The Economy of Love and Fear: A Preface to Grants Economics.* By Harry G. Johnson 1055
Cagan, Phillip. *The Channels of Monetary Effects on Interest Rates.* By William Poole 665
Cahill, Harry A. *The China Trade and U.S. Tariffs.* By Robert F. Dernberger 1056
Caves, Richard E., and Jones, Ronald W. *World Trade and Payments: An Introduction.* By Anne O. Krueger 1058
Downs, Anthony. *Opening Up the Suburbs.* By Edwin S. Mills. 1317
Ehrenberg, Ronald G. *The Demand for State and Local Government Employees.* By Richard W. Parks 1317
Gordon, David M. *Theories of Poverty and Unemployment.* By Sherwin Rosen 437
Kirsch, Leonard J. *Soviet Wages: Changes in Structure and Administration since 1956.* By Murray Yanowitch 439
Mundell, Robert A. *Monetary Theory: Inflation, Interest, and Growth in the World Economy.* By Don Patinkin 217
Rader, Trout. *Theory of General Economic Equilibrium.* By James Quirk . . 220
Rosenberg, Nathan. *Technology and American Economic Growth.* By Lars G. Sandberg 221
Webber, Michael J. *Impact of Uncertainty on Location.* By Jacques Thisse. . 440
Weber, Arnold R. *In Pursuit of Price Stability: The Wage-Price Freeze of 1971.* By William Poole 1060
Zaleski, Eugene. *Planning for Economic Growth in the Soviet Union, 1918–1932.* By R. W. Davies 222

Puzzles and Problems 215

REVIEW OF ECONOMIC STUDIES

Accession Number
169955
Date 26.8.96

DR. ZAKIR HUSAIN LIBRARY
PERIODICAL SECTION
J.M.I., N. DELHI

Edited by G M Heal and D F Hendry

The Review of Economic Studies was founded in 1933 by a group of young British and American economists. It is published by the Society for Economic Analysis Board whose object is to encourage research in theoretical and applied economics, particularly by young economists. It draws its editors and referees from both sides of the Atlantic. Articles on theoretical and applied economics are published with an emphasis on the development and testing of theory. It has become a principal Journal in the field of mathematical economics without being exclusively devoted to work of this kind. In particular there has recently been a considerable increase in the emphasis placed on econometric articles.

Reviews of books are not published.

Published quarterly in January, April, July and October
Volume 42, 1975

Annual subscription	£9.00	USA $27.00
Single numbers	£2.50	USA $7.50
Student rate	£1.75	USA $4.00
Index to Volumes 1 – 25 (1933-58)	£2.00	USA $6.00

Private individual subscribers who send their subscriptions to Longman Group together with a signed statement that the subscription is for their private use can, pay a reduced rate of £7.00 or USA S21.00.

Longman Group Limited, Journals Division, 43/45 Annandale Street,
Edinburgh EH7 4AT, Scotland.

Longman
1724 - 1974

Journal of Political Economy

acknowledges the assistance of:

Douglas Adie
Dennis Aigner
William Allen
Kenneth Arrow
Elizabeth E. Bailey
Martin J. Bailey
Martin N. Baily
Stephen Barro
Yoram Barzel
William Baumol
Yoram Ben-Porath
George Benston
Alan S. Blinder
Michael Boskin
David F. Bradford
William Branson
O. H. Brownlee
Neil Bruce
Michael E. Canes
John A. Carlson
Barry Chiswick
A. W. Coats
David Davies
Frank De Leeuw
Harold Demsetz
Arthur De Vany
William Dewald
Donald Dewey
Peter A. Diamond
Billy J. Eatherly
Ronald Ehrenberg
Bryan Ellickson
Robert Evenson
Allan Feldman
James Ferguson
Stanley Fischer
Frank Flatters
John E. Floyd
H. E Frech
Richard Freeman
Alan Frieden
James W. Friedman
Victor Fuchs
Mason Gaffney
Arthur Gandolfi
Allan Gibbard
Claudia Goldin
Robert J. Gordon
J. R. Gould
M. L. Greenhut
Reuben Gronau
Herschel Grossman
Michael Grossman
Robert Hall
Daniel Hamermesh
Giora Hanoch
W. Lee Hansen
James Hanson
Milton Harris
Jack Hirshleifer
Peter Howitt

Rodney Jacobs
Robert Jones
Stephan F. Kaliski
Morton I. Kamien
Edward Kane
John Kareken
Arnold Katz
Donald Katzner
George Kaufman
Harry Kelejian
John Kennan
Mohsin Khan
Richard Kihlstrom
Alvin K. Klevorick
Neil Komesar
Anthony Koo
Pentti Kouri
Gerald Kramer
Melvyn Krauss
Anne O. Krueger
Howard Kunreuther
Michael Landsberger
Lawrence J. Lau
Robert Laurent
Richard Layard
John Ledgard
Tong Hun Lee
Harvey Leibenstein
Arleen Leibowitz
Axel Leijonhufvud
David Levhari
Michael C. Lovell
John McCall
Huston McCulloch
Wyatt Mankin
Edwin Mansfield
Jorge Marquez
A. James Meigs
Peter Mieszkowski
Jacob Mincer
Edward Mitchell
J. M. Montias
Claude Montmarquette
Thomas Moore
M. Morishima
J. R. Moroney
Dennis C. Mueller
Patricia Munch
Michael Mussa
Richard F. Muth
Keizo Nagatani
Phillip Nelson
Marc Nerlove
Joseph P. Newhouse
Jeffrey Nugent
Lawrence Officer
Randall Olsen
Lawrence Olson
June O'Neill
Daniel Orr
Joseph Ostroy

Bruce Owen
John Owen
Robert Parks
Donald O. Parsons
Mark Pauly
Alan Peacock
John Marshall Peterson
Edmund S. Phelps
Solomon Polachek
Mitchell Polinsky
Uri Possen
George Psacharopoulos
Assaf Razin
Marcel Richter
Paul Craig Roberts
Carlos Rodriguez
Sherwin Rosen
Michael Rothschild
Larry E. Ruff
Roy Ruffin
Walter S. Salant
Steven C. Salop
Paul Samuelson
Agnar Sandmo
Anthony Santomero
Thomas Sargent
Thomas Saving
Richard Schuler
T. Paul Schultz
David Schwartzman
Robert Honey Scott
John Siegfried
Vernon Smith
Eugene Smolensky
Lee Soderstrom
Michael Spence
Case Sprenkle
Frank Stafford
Ross Starr
Frank G. Steindl
Peter L. Swan
T. Nicolas Tideman
Arthur B. Treadway
Jack E. Triplett
Chris Veendorp
Terence Wales
Donald A. Walker
David Wall
Alan Walters
Benjamin Ward
John Weicher
Leonard W. Weiss
Finis Welch
Andrew B. Whinston
Oliver Williamson
Charles Wilson
Gordon Winston
Dan Wisecarver
Gavin Wright
Shih-Yen Wu
Eugene Zajac

The Pure Theory of Money III

The whole civilized world is now eager to know whether in the future the high cost of living is to advance further, recede, or remain stationary. Opinions are plentiful but data supporting them are few. Even the best forecasts I have seen appear to be based on a very incomplete comprehension of the problem. Many conceive it as a problem of ordinary supply and demand and discuss the general *price level* as they would discuss the *price* of wheat or any other commodity, overlooking the fact that the causes affecting price levels are as distinct from those affecting an individual price as the causes affecting the tides are distinct from those affecting an individual wave.

The problem of the cost of living is primarily a problem of the general level of prices, although it is also partly a problem of special prices. If the phrase "cost of living" were used to refer to the general level of all prices the problem would be purely of the former type and not at all of the latter. But this phrase is usually taken as referring only to special groups of commodities, mostly foods, and only for the retail prices of these commodities. . . .

If I am correct in my philosophy of price levels . . . the general level of prices in the world is determined by the other . . . magnitudes which are joined with it in "the equation of exchange," [M, money in circulation; V, its velocity; and T, the volume of trade.] No other considerations whatever are relevant—trusts, tariffs, trade unions, shorter hours, limitation of output of labor, exhaustion of the soil, concentration of population in cities, middlemen, advertising, overcapitalization, restrictive legislation, cold storage, pure food legislation, sanitary legislation, food adulteration, the "individual package," extravagance, world armaments, wars, old-age pensions, unemployment, etc.—except so far as these factors affect one or more of the . . . factors in the equation of exchange which alone can act on the general level of prices. . .

The proper mode of procedure is therefore to make a forecast as to these . . . factors. Thus far only one of them, the volume of money in circulation, has received much attention, and even the discussion of this factor has not preceeded along the lines proper for a correct forecast. . . .

The problem . . . is to forecast the future world stock of money. But it is more than this. Even if we forecast the world's stock of money, we must still take account of the other . . . price-influencing factors in the equation of exchange. . . .

[Irving Fisher, "Will the Present Upward Trend of World Prices Continue?" *American Economic Review*, September 1912]